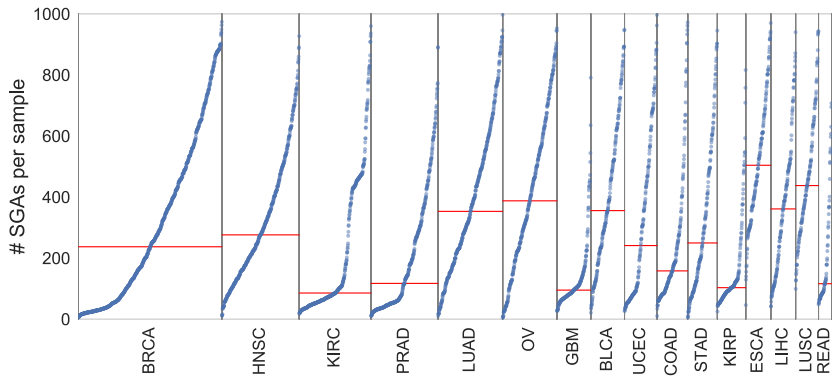

Fig. S1. Somatic genomic alteration (SGA) distribution of different cancer types in the TCGA dataset. 16 cancer types are sorted in descending order of sample size. Tumor samples of each cancer type are further sorted in ascending order of SGA numbers. Each red line shows the median number of SGAs in the tumors of specific cancer type.

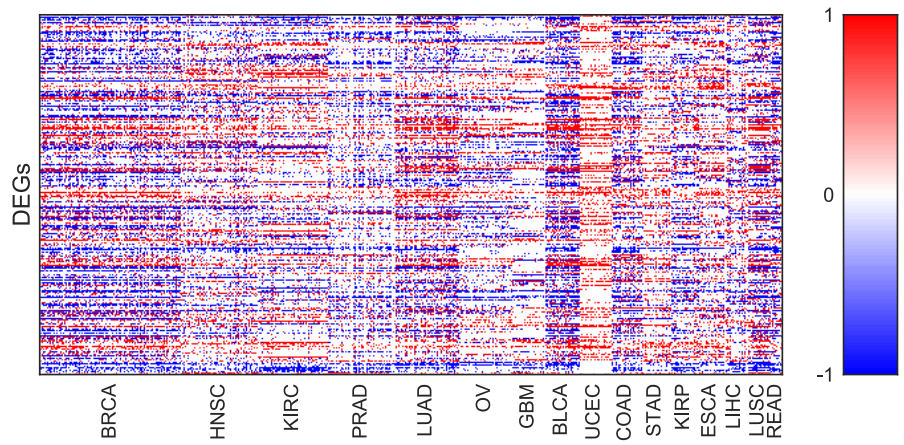

Fig. S2. Differentially expressed gene (DEG) distribution of different cancer types in the TCGA dataset. 16 cancer types are sorted in descending order of sample numbers. Each column shows the gene expression profiles of specific tumor size. “1”, “-1” and “0” mean over-, under- and unchanged expression 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 $\mathbf{e}_s$ to GIT model.

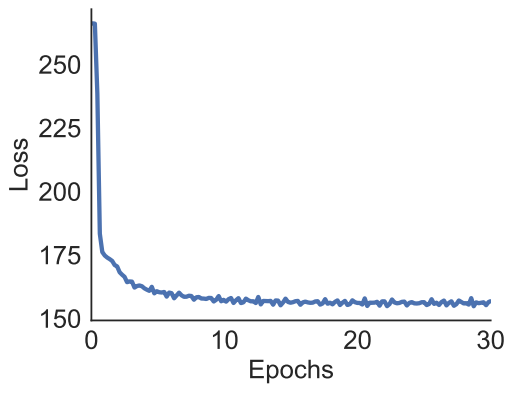

Fig. S3. Negative sampling loss of Gene2Vec algorithm during the pre-training of gene embeddings. Skip gram method with negative sampling loss is utilized to pre-train the gene embeddings which can reflect the co-occurrence pattern of somatic genomic alterations (SGAs) in tumors. Hyperparameters including 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.

Data: Genomic alterations in each tumor: $\mathcal{T} = \left\{T_i = \{g_{i1}, g_{i2}, ..., g_{im(i)}\}\right\}_{i=1,2,...,N}$.
Result: Pretrained gene embedding of each gene: $\mathcal{E} = \{\mathbf{e}_g \in \mathbb{R}^n\}_{g \in \mathcal{G}}$. Context gene embeddings: $\mathcal{V} = \{\mathbf{v}_g \in \mathbb{R}^n\}_{g \in \mathcal{G}}$.
$f(g) \leftarrow \frac{1}{Z} \sum_{i=1}^{N} \mathbb{1}(g \in T_i),\ g \in \mathcal{G}$; `// Frequency of each gene`
$f_n(g) \leftarrow \frac{1}{Z_n} f(g)^{3/4},\ g \in \mathcal{G}$; `// Normalized sub-sampled frequency`
$\mathbf{e}_g \sim U\left(-\frac{0.5}{n}, \frac{0.5}{n}\right)^n,\ \mathbf{v}_g \leftarrow 0^n,\ g \in \mathcal{G}$; `// Initialize gene embeddings and context embeddings`
while *not converges* **do**
  $l \leftarrow 0$; `// Total loss of a mini-batch samples`
  for $b = 1, 2, ...,$ *batch_size* **do**
    $g \sim f$; `// Sample a gene`
    $g_c \sim$ Context$(g\ ;\ \mathcal{T})$; `// Sample a context gene`
    $g_{nr} \sim f_n,\ r = 1, 2, ..., R$; `// Sample negative context genes`
    $l \leftarrow l +$ NSLoss $\left(g, g_c, \{g_{nr}\}_{r=1}^{R}\ ;\ \mathcal{E}, \mathcal{V}\right)$; `// Update loss of negative sampling`
  end
  $(\mathcal{E}, \mathcal{V}) \leftarrow (\mathcal{E}, \mathcal{V}) - \eta \cdot \frac{\partial l}{\partial (\mathcal{E}, \mathcal{V})}$; `// Update gene and context embeddings using gradient descent`
end
Function `Context(`$g\ ;\ \mathcal{T}$`)`
  $P_c \leftarrow U\left(\{g_c \mid g_c \in T_i, g \in T_i\}_{i=1,2,...,N}\right)$; `// Uniform distribution over sequence of adjacent mutations`
return P_c
Function `NSLoss(`$g, g_c, \{g_{nr}\}_{r=1}^{R}\ ;\ \mathcal{E}, \mathcal{V}$`)`
  $l \leftarrow \log \sigma\left(\mathbf{e}_g^\intercal \mathbf{v}_{g_c}\right) + \sum_{r=1}^{R} \log \sigma\left(-\mathbf{e}_g^\intercal \mathbf{v}_{g_{nr}}\right)$; `// 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 probability of co-occurred SGAs c given a 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 $f(g)^{3/4}$. In practice, the step size η in mini-batch gradient descent is decayed after training for every epoch to converge fast and prevent overfitting. Note that $\mathcal{E}$ 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. **Gene2Vec**: Stochastic gradient descent (SGD) is used following word2vec (Mikolov *et al*., 2013). The step size is decayed by 0.78 after training for every epoch. The ratio of negative sample size to positive sample size is set to be 5. **Lasso**: It requires more epochs and larger step size to converge due to its large number of parameters. **MLPs**: Dropout is applied at each hidden layer to reduce overfitting. It is not shown in the structure below due to space limit. **GIT**: There are additional structure parameters for the attention module. The dimension of attention parameter β_j is 400, and the number of heads h is set to be 128 (j=1,2,...,128). **Notations**: "ns" means "negative sampling", "emb" means "embedding layer", "σ" means "sigmoid activation function", "fc" means "fully connected layer", "relu" means "ReLU activation function", "attn" means "attention mechanism".

Methods	Model structure	Optimizer	Step size	Epochs	Batch size	Dropout	λ_1	λ_2
Gene2Vec	$19.8\text{k} \xrightarrow{\text{ns+emb+}\sigma} 19.8\text{k}$	SGD	0.128	30	64	–	–	–
Lasso	$19.8\text{k} \xrightarrow{\text{fc+}\sigma} 2.2\text{k}$	Adam	3e-3	42	16	–	1.0	–
1 layer MLP	$19.8\text{k} \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+}\sigma} 2.2\text{k}$	Adam	3e-4	31	16	0.5	–	1e-9
2 layer MLP	$19.8\text{k} \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+}\sigma} 2.2\text{k}$	Adam	1e-4	31	16	0.5	–	1e-5
3 layer MLP	$19.8\text{k} \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+relu}} 512 \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+}\sigma} 2.2\text{k}$	Adam	1e-4	31	16	0.5	–	1e-5
GIT	$19.8\text{k} \xrightarrow{\text{emb+attn+relu}} 512 \xrightarrow{\text{fc+relu}} 1024 \xrightarrow{\text{fc+}\sigma} 2.2\text{k}$	Adam	1e-4	31	16	0.5	–	1e-5

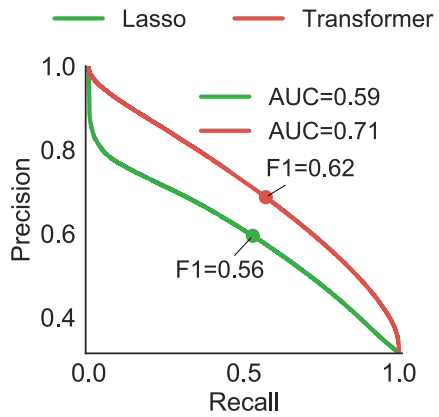

Fig. S4. Precision-recall (PR) curve of Lasso and GIT in predicting differentially expressed genes (DEGs) from somatic genomic alterations (SGAs). Not only does GIT get larger F1 score (F1 = 0.62) than conventional Lasso method (F1 = 0.56), GIT also has a higher PR curve (AUC = 0.71) than Lasso (AUC = 0.59), validating the superiority of our method.

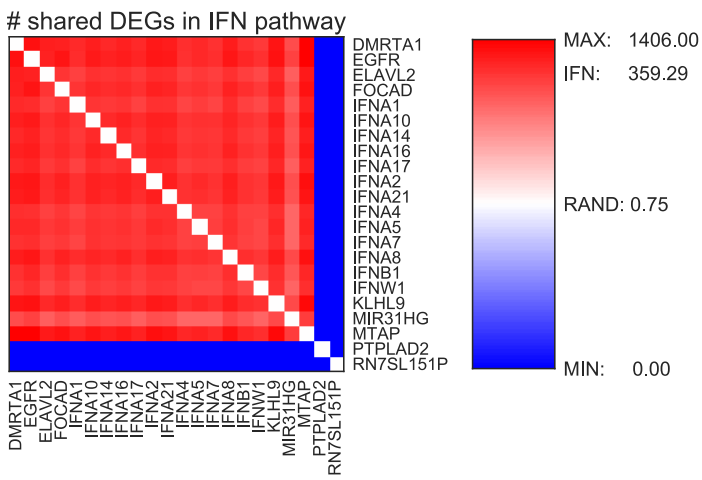

Fig. S5. Number of shared differentially expressed genes (DEGs) caused by somatic genomic alterations (SGAs) in the "IFN pathway". Based on the results of causal relationship of SGAs and DEGs from Tumor-specific Causal Inference (Cai *et al.*, 2018), we calculate the number of shared caused DEGs of any 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.

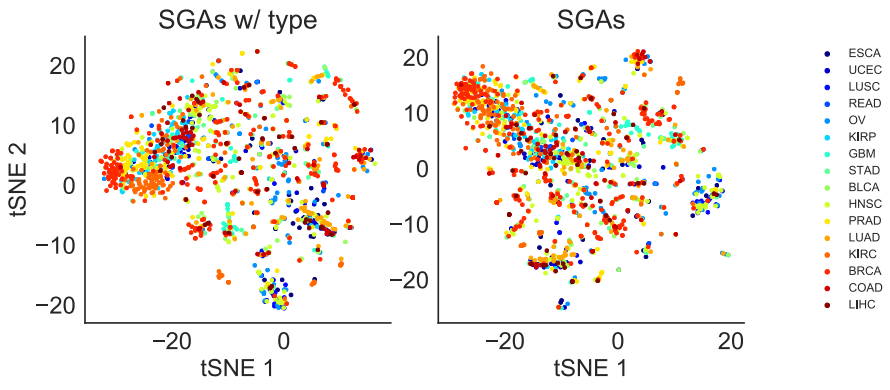

Fig. S6. t-SNE visualization of tumors represented as the raw sparse SGA data. Left panel shows the case when one-hot cancer type vector is concatenated with binary SGA vector, while right panels is simply the case of binary SGA vector. Opposed to represented as compact tumor embeddings in Figure 5**a,b**, 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.

Table S2. Groups of genes using *k*-means clustering in the gene embedding space. Genes that are altered in at least 150 out of 4,468 tumors are clustered to filter out genes whose gene embeddings that may not be learned well. 1,474 candidate genes in total are finally analyzed.

Cluster ID	Set of genes
1	*ABRA*, *ANGPT1*, *C8orf56*, *CALB1*, *CCNE2*, *CDH17*, *COL14A1*, *CPQ*, *DCSTAMP*, *DECR1*, *DPY19L4*, *EBAG9*, *EIF3E*, *EMC2*, *ESRP1*, *EYA1*, *FAM92A1*, *FSBP*, *FZD6*, *GEM*, *INTS8*, *KIAA1429*, *LRP12*, *MCM4*, *MIR2053*, *MIR3150B*, *MIR548A3*, *MTDH*, *MTERFD1*, *NBN*, *NECAB1*, *ODF1*, *OSGIN2*, *OTUD6B*, *OXR1*, *POP1*, *PTDSS1*, *RAD54B*, *RBM12B*, *RIPK2*, *RN7SKP231*, *RN7SL777P*, *RNU6ATAC8P*, *RUNX1T1*, *SDC2*, *SLC26A7*, *SPAG1*, *STK3*, *TMEM55A*, *TMEM64*, *TMEM67*, *TMEM74*, *TRHR*, *TSPYL5*, *UBR5*, *UQCRB*, *ZFPM2*
2	*ARFGEF2*, *ASXL1*, *ATP9A*, *BCAS1*, *CASS4*, *CDH26*, *CDH4*, *CHD6*, *CHRNA4*, *COL20A1*, *CTCFL*, *CYP24A1*, *DIDO1*, *DOK5*, *FAM65C*, *HELZ2*, *KCNB1*, *KCNQ2*, *LAMA5*, *MYH7B*, *MYT1*, *NPBWR2*, *NTSR1*, *OGFR*, *OPRL1*, *OSBPL2*, *PCK1*, *PCMTD2*, *PHACTR3*, *PREX1*, *PRPF6*, *PTPRT*, *RAE1*, *RTEL1*, *SALL4*, *SLCO4A1*, *SRMS*, *TAF4*, *TSHZ2*, *ZBTB46*, *ZFP64*, *ZGPAT*, *ZMYND8*, *ZNF217*, *ZNF512B*, *ZNF831*, *ZNFX1*
3	*ACTN2*, *AHCTF1*, *AKT3*, *ARID4B*, *ASPM*, *ATP2B4*, *C1orf101*, *CACNA1E*, *CDC42BPA*, *CEP170*, *CFH*, *CHML*, *CHRM3*, *CNST*, *CNTN2*, *CR1*, *CR2*, *CRB1*, *DISC1*, *DISP1*, *EXO1*, *HEATR1*, *HNRNPU*, *IRF2BP2*, *ITPKB*, *KDM5B*, *KIF14*, *KIF26B*, *LRRN2*, *LYST*, *MTR*, *NID1*, *OBSCN*, *OR11L1*, *OR13G1*, *OR14A16*, *OR14C36*, *OR14I1*, *OR1C1*, *OR2AK2*, *OR2B11*, *OR2C3*, *OR2G2*, *OR2G3*, *OR2G6*, *OR2L13*, *OR2L2*, *OR2L3*, *OR2L8*, *OR2M2*, *OR2M3*, *OR2M4*, *OR2M5*, *OR2M7*, *OR2T1*, *OR2T11*, *OR2T12*, *OR2T2*, *OR2T27*, *OR2T3*, *OR2T33*, *OR2T34*, *OR2T4*, *OR2T6*, *OR2T8*, *OR2W3*, *OR2W5*, *OR6F1*, *PCNXL2*, *PGBD2*, *PLEKHA6*, *PPP1R15B*, *PROX1*, *PRRC2C*, *PTPRC*, *RGS7*, *RYR2*, *SDCCAG8*, *SIPA1L2*, *TARBP1*, *TRIM58*, *URB2*, *USH2A*, *WDR64*, *ZC3H11A*, *ZNF496*, *ZP4*
4	*ABCB1*, *AC008103.5*, *ADAMTS18*, *ADAMTS20*, *AKAP11*, *AKAP9*, *ALPK2*, *ANK3*, *ANKRD11*, *ASCC3*, *BAGE2*, *BRCA2*, *BSN*, *CDH13*, *CDH7*, *CELSR3*, *CHD3*, *CMYA5*, *CPM*, *DGKB*, *DIAPH3*, *DNAH1*, *DOCK3*, *DSEL*, *EPHA7*, *EPHB1*, *EYS*, *FBN2*, *FLT4*, *GPR112*, *GRIK2*, *HECW1*, *HYDIN*, *IL1RAPL1*, *KIAA1109*, *LAMA4*, *LRP2*, *LRRC7*, *LRRK2*, *MAGEC1*, *MDM2*, *MDN1*, *MLL3*, *MST1P2*, *MXRA5*, *MYH1*, *MYH8*, *MYO16*, *NCOA6*, *NEB*, *NES*, *NLGN4X*, *NOTCH3*, *ODZ1*, *PCDH10*, *PCDH15*, *PEG3*, *PKD1*, *PKD1L2*, *PKHD1*, *PTPRB*, *RALGAPB*, *RANBP2*, *SALL3*, *SAMD9*, *SCN10A*, *SCN11A*, *SDK2*, *SIPA1L3*, *SLC12A5*, *SMAD4*, *SMG1*, *SRCAP*, *STAB1*, *TENM3*, *THSD7A*, *THSD7B*, *TNR*, *TNRC18*, *TNXB*, *TRANK1*, *TRHDE*, *TRPC4*, *TRRAP*, *TSSC2*, *ZNF292*, *ZNF407*, *ZNF479*
5	*AADAC*, *ACAP2*, *ACTL6A*, *ANKRD18DP*, *APOD*, *ARHGEF26*, *DLG1*, *GMPS*, *GNB4*, *HRASLS*, *KCNMB2*, *KIAA0226*, *LEPREL1*, *LINC00969*, *LRRC15*, *MECOM*, *MED12L*, *MFN1*, *MIR4797*, *MIR569*, *MIR944*, *MME*, *MUC4*, *NAALADL2*, *NLGN1*, *PEX5L*, *PLD1*, *PLSCR4*, *RN7SKP234*, *RN7SKP40*, *RNA5SP149*, *RNF168*, *RNY5P3*, *SDHAP1*, *SLC2A2*, *SLC7A14*, *TBL1XR1*, *TM4SF19*, *TMEM44*, *TNIK*, *TP63*, *TPRG1*, *TTC14*, *U2SURP*, *USP13*, *XRN1*, *ZBTB38*
6	*ABCA10*, *AHNAK*, *ANK2*, *FAM47C*, *MUC2*, *PIK3R1*, *PTPRD*, *PTPRZ1*, *SETD2*, *SPOP*, *TEX15*
7	*ADAMTS16*, *ADCY2*, *AHRR*, *ANKH*, *BRD9*, *C5orf38*, *C5orf49*, *CCDC127*, *CCT5*, *CDH12*, *CLPTM1L*, *CMBL*, *CT49*, *DNAH5*, *FAM105A*, *FAM105B*, *FAM134B*, *FAM173B*, *FASTKD3*, *FBXL7*, *IRX1*, *IRX2*, *IRX4*, *KIAA0947*, *LPCAT1*, *LRRC14B*, *MIR4277*, *MIR4457*, *MIR4458*, *MIR4635*, *MIR4636*, *MIR4637*, *MTRR*, *NKD2*, *NSUN2*, *PAPD7*, *PDCD6*, *PLEKHG4B*, *PRDM9*, *RN7SKP73*, *RN7SL58P*, *RNA5SP177*, *RNA5SP180*, *SDHA*, *SDHAP3*, *SEMA5A*, *SLC12A7*, *SLC6A18*, *SLC6A19*, *SLC6A3*, *SLC9A3*, *SRD5A1*, *TAS2R1*, *TERT*, *TRIO*, *TRIP13*, *ZDHHC11*, *ZNF622*
8	*ACAN*, *ALMS1*, *ANKRD20A8P*, *ANKRD30A*, *ATN1*, *BCOR*, *CD163L1*, *CDH10*, *CDH19*, *CHD4*, *CNTNAP5*, *COL6A3*, *CSMD2*, *DCC*, *DCHS2*, *DNAH17*, *DUSP27*, *FCGBP*, *FRG1*, *GRIA1*, *LAMA1*, *LAMA3*, *LRP1B*, *MUC6*, *MYLK*, *NFE2L2*, *NOTCH1*, *PDE3A*, *POLQ*, *POM121L12*, *POTEC*, *SNHG14*, *SYNE1*, *VWF*, *WDR17*, *WNK1*, *ZNF318*, *ZNF521*, *ZNF804B*
9	*ANGPT2*, *BLK*, *C8orf12*, *CTSB*, *DEFA5*, *DEFB1*, *ERI1*, *ERICH1*, *FAM66B*, *FAM66D*, *FAM86B1*, *FAM86B2*, *FDFT1*, *KBTBD11*, *LONRF1*, *MCPH1*, *MFHAS1*, *MIR5692A2*, *MTMR9*, *MYOM2*, *NEIL2*, *PINX1*, *PRSS55*, *RN7SL872P*, *RPL23AP53*, *SGK223*, *SLC35G5*, *SOX7*, *TDH*, *TNKS*, *USP17L2*, *XKR5*, *XKR6*, *ZNF596*, *ZNF705D*
10	*ACY3*, *ALDH3B2*, *CPT1A*, *FAM86C2P*, *GAL*, *KDM2A*, *LRP5*, *MALAT1*, *MRPL21*, *MTL5*, *NDUFV1*, *PITPNM1*, *PPP6R3*, *SUV420H1*, *TBX10*
11	*ADCK5*, *ADCY8*, *AGO2*, *ARC*, *ARHGAP39*, *ASAP1*, *BAI1*, *C8orf33*, *COL22A1*, *CPSF1*, *CYC1*, *CYP11B1*, *CYP11B2*, *DENND3*, *EEF1D*, *EPPK1*, *FAM135B*, *FAM203A*, *GLI4*, *GPIHBP1*, *KCNK9*, *KCNQ3*, *KHDRBS3*, *MAFA*, *MROH5*, *NAPRT1*, *OPLAH*, *PHF20L1*, *PLEC*, *RECQL4*, *RHPN1*, *SCRIB*, *SLA*, *SLC45A4*, *SPATC1*, *ST3GAL1*, *TG*, *TMEM71*, *TONSL*, *TRAPPC9*, *ZC3H3*, *ZFAT*, *ZNF16*, *ZNF251*, *ZNF252P*, *ZNF34*, *ZNF623*
12	*ARFGEF1*, *CA1*, *CHCHD7*, *CRISPLD1*, *EXT1*, *FAM110B*, *IMPAD1*, *KCNB2*, *LYN*, *MOS*, *MRPL15*, *MYBL1*, *PENK*, *PKHD1L1*, *PRKDC*, *RALYL*, *RGS20*, *SOX17*, *ST18*, *STAU2*, *STMN2*, *TCEA1*, *TERF1*, *TGS1*, *TRAM1*, *TRPA1*, *UBE2W*, *XKR4*, *ZFHX4*
13	*ADAM2*, *ADAM32*, *ADAM9*, *AGPAT6*, *ANK1*, *AP3M2*, *ASH2L*, *DDHD2*, *FGFR1*, *FNTA*, *HGSNAT*, *HOOK3*, *HTRA4*, *IKBKB*, *KAT6A*, *LETM2*, *MIR486*, *PLAT*, *PLEKHA2*, *POLB*, *POTEA*, *PPAPDC1B*, *RN7SKP41*, *RN7SL149P*, *RNF170*, *SGK196*, *SLC20A2*, *STAR*, *TACC1*, *TM2D2*, *WHSC1L1*
14	*DMRTA1*, *EGFR*, *ELAVL2*, *FOCAD*, *IFNA1*, *IFNA10*, *IFNA14*, *IFNA16*, *IFNA17*, *IFNA2*, *IFNA21*, *IFNA4*, *IFNA5*, *IFNA7*, *IFNA8*, *IFNB1*, *IFNW1*, *KLHL9*, *MIR31HG*, *MTAP*, *PTPLAD2*, *RN7SL151P*
15	*CDK12*, *GRB7*, *IKZF3*, *NEUROD2*, *PGAP3*, *PPP1R1B*, *STARD3*, *ZPBP2*
16	*ADAMTS12*, *AGXT2*, *AMACR*, *C1QTNF3*, *C5orf22*, *C5orf34*, *C5orf42*, *C6*, *C9*, *CARD6*, *CDH6*, *DAB2*, *DROSHA*, *EGFLAM*, *FYB*, *GHR*, *GUSBP1*, *IL7R*, *LIFR*, *LMBRD2*, *MROH2B*, *MTMR12*, *NIPBL*, *NPR3*, *NUP155*, *OSMR*, *OXCT1*, *PAIP1*, *PLCXD3*, *PRKAA1*, *PRLR*, *PTGER4*, *RAI14*, *RANBP3L*, *RICTOR*, *RXFP3*, *SAMD9L*, *SKP2*, *SLC1A3*, *SLC45A2*, *SPEF2*, *TARS*, *TTC33*, *UGT3A1*, *UGT3A2*, *WDR70*, *ZFR*, *ZNF131*
17	*AAMDC*, *ALG8*, *ARAP1*, *ARHGEF17*, *B3GNT6*, *C11orf30*, *C2CD3*, *CAPN5*, *DHCR7*, *FAM86C1*, *GAB2*, *GDPD4*, *INPPL1*, *INTS4*, *KCTD14*, *KCTD21*, *LRRC32*, *MYO7A*, *NADSYN1*, *NUMA1*, *PAK1*, *PDE2A*, *RSF1*, *TENM4*, *UNC93B6*, *USP35*
18	*ABCC5*, *ABCF3*, *ADIPOQ*, *AHSG*, *AP2M1*, *ATP11B*, *ATP13A3*, *ATP13A4*, *ATP13A5*, *BCL6*, *BDH1*, *C3orf65*, *C3orf70*, *CCDC39*, *CCDC50*, *CHRD*, *CLCN2*, *CLDN11*, *CPN2*, *DCUN1D1*, *DGKG*, *DNAJB11*, *DNAJC19*, *DVL3*, *ECE2*, *ECT2*, *EHHADH*, *EIF2B5*, *EIF4A2*, *EIF4G1*, *EPHB3*, *ETV5*, *FAM131A*, *FAM157A*, *FAM43A*, *FETUB*, *FGF12*, *FNDC3B*, *FXR1*, *GP5*, *HRG*, *HTR3C*, *HTR3D*, *HTR3E*, *IL1RAP*, *KLHL24*, *KLHL6*, *KNG1*, *LAMP3*, *LINC00884*, *LMLN*, *LPP*, *LRCH3*, *LRRC31*, *LSG1*, *MAGEF1*, *MAP3K13*, *MASP1*, *MB21D2*, *MCCC1*, *MCF2L2*, *MFI2*, *MIR28*, *MUC20*, *NCEH1*, *OPA1*, *PAK2*, *PHC3*, *PIGZ*, *PP13439*, *PRKCI*, *RFC4*, *RN7SKP296*, *RN7SL141P*, *RN7SL229P*, *RN7SL447P*, *RN7SL486P*, *RN7SL637P*, *RTP1*, *RTP2*, *SAMD7*, *SENP2*, *SERPINI2*, *SKIL*, *SPATA16*, *TFRC*, *TMEM207*, *TMEM212*, *TNK2*, *TRA2B*, *VPS8*, *YEATS2*
19	*BAP1*, *MUC17*, *MYCBP2*, *PTEN*, *UNC13C*, *ZFHX3*
20	*PBRM1*, *VHL*
21	*ANO1*, *CTTN*, *FADD*, *IGHMBP2*, *MIR3664*, *MIR548K*, *MRGPRD*, *MRGPRF*, *MYEOV*, *PPFIA1*, *SHANK2*
22	*ADAMTS4*, *ADAMTSL4*, *ANKRD35*, *ARHGAP30*, *ARNT*, *ATAD1*, *BCL9*, *C1orf51*, *CERS2*, *CGN*, *CHD1L*, *DDR2*, *DENND4B*, *ECM1*, *F5*, *FAM63A*, *FLG*, *FMO5*, *GJA8*, *HORMAD1*, *INSRR*, *ITGA10*, *IVL*, *KLLN*, *KPRP*, *LY9*, *MTMR11*, *NBPF10*, *NBPF14*, *NBPF9*, *NOTCH2*, *OTUD7B*, *PDE4DIP*, *PI4KB*, *PIAS3*, *PIP5K1A*, *POGZ*, *POLR3C*, *PRPF3*, *PRUNE*, *RFX5*, *RPRD2*, *RPTN*, *SCNM1*, *SEC22B*, *SELENBP1*, *SEMA6C*, *SETDB1*, *SPTA1*, *SV2A*, *TARS2*, *TCHH*, *TCHHL1*, *TXNIP*, *ZNF687*
23	*ABCA9*, *ATXN1*, *BAI3*, *CELSR1*, *COL11A1*, *COL1A2*, *CTNNB1*, *DCHS1*, *EP300*, *EP400*, *FREM2*, *FRG1B*, *GATA3*, *HRNR*, *KBTBD6*, *KEAP1*, *MACF1*, *MUC16*, *NALCN*, *NCKAP5*, *NFATC2*, *NRXN1*, *NSD1*, *PCDH17*, *RCBTB2*, *RNF213*, *ROBO1*, *ROBO2*, *SSPO*, *SYCP2*, *TTI2*, *VCAN*, *ZNF814*
24	*APC*, *CCSER1*, *FBXW7*, *PDE4D*, *RB1*, *RYR1*, *WWOX*
25	*ACE*, *APOB*, *BEST3*, *BMP7*, *C1orf173*, *CD163*, *CDH23*, *CLTC*, *CNTNAP2*, *CUX1*, *DMD*, *DNAH7*, *DNAH8*, *DNM1P47*, *DOCK2*, *DST*, *F8*, *FAM208B*, *LAMA2*, *LRP1*, *MKI67*, *MLL2*, *MUC5B*, *NBEA*, *NF1*, *NUP107*, *SDK1*, *TAF1L*, *TPR*, *TTN*, *WASH3P*, *ZNF536*, *ZNF733P*, *ZNF99*
26	*CDH1*
27	*C9orf53*, *CDKN2A*, *CDKN2B*
28	*ADRA1A*, *CHRNA2*, *CLU*, *DOCK5*, *DPYSL2*, *DUSP4*, *EBF2*, *ELP3*, *EPHX2*, *ESCO2*, *EXTL3*, *FBXO16*, *FZD3*, *GSR*, *INTS9*, *KIF13B*, *LEPROTL1*, *MAK16*, *MIR548H4*, *NEFM*, *PBK*, *PNMA2*, *PNOC*, *PPP2CB*, *PPP2R2A*, *PTK2B*, *PURG*, *RBPMS*, *RN7SL781P*, *RNA5SP258*, *RNA5SP261*, *SCARA3*, *SCARA5*, *STMN4*, *TMEM66*, *TRIM35*, *TUBBP1*, *UBXN8*, *WRN*, *ZNF395*
29	*ANXA13*, *ATAD2*, *C8orf47*, *CNGB3*, *CSMD3*, *DCAF13*, *DPYS*, *EFR3A*, *ENPP2*, *FAM91A1*, *FER1L6*, *GDF6*, *GRHL2*, *HAS2*, *KCNS2*, *KCNV1*, *KIAA0196*, *KLHL38*, *LAPTM4B*, *MATN2*, *MIR1273A*, *MIR4471*, *MRPL13*, *MTBP*, *MTSS1*, *NACAP1*, *NIPAL2*, *NOV*, *PABPC1*, *POU5F1B*, *RGS22*, *RIMS2*, *RN7SKP153*, *RN7SL563P*, *RN7SL590P*, *RNF19A*, *RNY4P5*, *SAMD12*, *SLC30A8*, *SNTB1*, *SYBU*, *TAF2*, *TNFRSF11B*, *TRPS1*, *VPS13B*, *ZHX1*, *ZHX2*
30	*ABCA6*, *ABCB5*, *ABCC9*, *ADAMTS2*, *AHNAK2*, *ARAP3*, *ARID1A*, *ARID1B*, *BZRAP1*, *CACNA1D*, *CACNA1G*, *COL6A6*, *COL7A1*, *CREBBP*, *CUBN*, *CUL9*, *DNAH3*, *EPHA6*, *FAM194B*, *FAT3*, *FAT4*, *FLNA*, *FRAS1*, *GLI3*, *GOLGB1*, *GPR98*, *GRIN2A*, *GRM5*, *HDAC9*, *HERC1*, *HIVEP3*, *ITPR2*, *KMT2C*, *KMT2D*, *LAMB4*, *LPAR6*, *MAP1B*, *MGA*, *MYH15*, *MYOM1*, *NBPF1*, *NEFH*, *PCDHB12*, *PCLO*, *PDGFRA*, *PDZD2*, *PKD1L1*, *PLXNA4*, *PLXNB2*, *RELN*, *RNLS*, *RRN3P2*, *RTTN*, *SACS*, *SCN4A*, *SLC6A10P*, *SLIT3*, *SPHKAP*, *SRRM2*, *TSHZ3*, *USP9X*, *VPS13D*, *XIRP2*, *ZC3H13*
31	*ERBB2*, *GSDMB*, *KIF2B*, *MED24*, *PIK3CA*, *PSMD3*, *THRA*
32	*CNTNAP4*, *DNAH14*, *DNAH9*, *FLG2*, *FMN2*, *HMCN1*, *HUWE1*, *KIF21B*, *MAP3K1*, *MYH13*, *MYH2*, *NCOA3*, *NCOR1*, *USP34*
33	*ADAM28*, *ADAM7*, *ADAMDEC1*, *ARHGEF10*, *ASAH1*, *ATP6V1B2*, *BIN3*, *BMP1*, *C8orf58*, *CDCA2*, *CSGALNACT1*, *DLC1*, *DLGAP2*, *DMTN*, *DOK2*, *EGR3*, *ENTPD4*, *FAM160B2*, *FP15737*, *GFRA2*, *HR*, *INTS10*, *KIAA1967*, *LOXL2*, *LPL*, *LZTS1*, *MIR383*, *MSR1*, *MTMR7*, *MTUS1*, *NAT1*, *NAT2*, *NEFL*, *NUDT18*, *PCM1*, *PDGFRL*, *PDLIM2*, *PHYHIP*, *PIWIL2*, *PPP3CC*, *PSD3*, *RHOBTB2*, *RN7SL651P*, *SFTPC*, *SGCZ*, *SH2D4A*, *SLC18A1*, *SLC25A37*, *SLC39A14*, *SLC7A2*, *SORBS3*, *STC1*, *TNFRSF10A*, *TNFRSF10B*, *TUSC3*, *VPS37A*, *XPO7*
34	*ABCA13*, *ANKRD30BP2*, *ATP7B*, *BIRC6*, *BRAF*, *DNAH11*, *DNAH2*, *DSCAM*, *FAT2*, *FBN3*, *FLNC*, *GNAS*, *HSD17B7P2*, *KRAS*, *LL22NC03-80A10.6*, *MYH4*, *NAV3*, *RYR3*, *SCN5A*, *SVEP1*, *TUBB8P7*, *ZAN*, *ZNF208*
35	*ANKHD1*, *ATM*, *ATP10B*, *BCAS3*, *BPTF*, *BRIP1*, *CACNA2D4*, *COL12A1*, *CPAMD8*, *CROCCP2*, *DNAH10*, *FAT1*, *GPR179*, *GRM6*, *HEATR6*, *HELZ*, *HERC2*, *KDM3B*, *KIF4B*, *MED1*, *MED13*, *MGAM*, *MRC2*, *MYH6*, *MYH7*, *MYO3A*, *PCDHA1*, *PCDHA10*, *PCDHA2*, *PCDHA7*, *PCDHAC2*, *PCDHB10*, *PCDHB6*, *PCDHB7*, *PCDHB8*, *PCDHGC5*, *PPM1E*, *RANBP17*, *RNF43*, *SPTBN2*, *SYNE2*, *TANC2*, *TBC1D3P2*, *TLK2*, *UQCRFS1*, *USP32*
36	*ADCY10*, *ASH1L*, *ASTN1*, *ATP1A2*, *ATP1A4*, *ATP8B2*, *CACNA1S*, *CENPF*, *CEP350*, *COPA*, *CRNN*, *EPRS*, *FAM5B*, *FAM5C*, *GON4L*, *IGFN1*, *IGSF9*, *KCNH1*, *KCNN3*, *KCNT2*, *KIAA0907*, *LAMC1*, *MIA3*, *NAV1*, *NFASC*, *NLRP3*, *NUP210L*, *PAPPA2*, *PIK3C2B*, *PLD5*, *PLXNA2*, *PRG4*, *PTPN14*, *SMG7*, *TNN*
37	*C19orf12*, *CACNA1C*, *CCNE1*, *IQSEC3*, *PLEKHF1*, *POP4*, *TP53*, *URI1*, *VSTM2B*
38	*ADHFE1*, *ARMC1*, *ASPH*, *ATP6V0D2*, *ATP6V1H*, *BHLHE22*, *C8orf34*, *C8orf44*, *C8orf46*, *CA2*, *CA3*, *CA8*, *CASC9*, *CHD7*, *CLVS1*, *CNBD1*, *COPS5*, *CPA6*, *CPNE3*, *CSPP1*, *CYP7A1*, *CYP7B1*, *DCAF4L2*, *DNAJC5B*, *E2F5*, *EFCAB1*, *FABP12*, *FAM150A*, *GGH*, *HEY1*, *HNF4G*, *IL7*, *JPH1*, *LINC00966*, *LRRCC1*, *LY96*, *LYPLA1*, *MCMDC2*, *MMP16*, *MRPS28*, *MSC*, *MTFR1*, *NCOA2*, *NKAIN3*, *NPBWR1*, *NSMAF*, *OPRK1*, *PAG1*, *PCMTD1*, *PDE7A*, *PEX2*, *PI15*, *PLAG1*, *PRDM14*, *PREX2*, *PSKH2*, *PXDNL*, *RAB2A*, *RB1CC1*, *REXO1L1*, *RMDN1*, *RN7SKP135*, *RN7SL107P*, *RN7SL308P*, *RNA5SP267*, *RNA5SP268*, *RNA5SP269*, *RNA5SP271*, *RNA5SP272*, *RRS1*, *SBSPON*, *SDCBP*, *SDR16C5*, *SDR16C6P*, *SGAgroup.85*, *SGK3*, *SLC10A5*, *SLC7A13*, *SLCO5A1*, *SNAI2*, *SNTG1*, *SNX16*, *SPIDR*, *SULF1*, *TOX*, *TRIM55*, *TTPA*, *UBXN2B*, *VCP1P1*, *WWP1*, *XKR9*, *YTHDF3*, *ZBTB10*, *ZNF704*
39	*C7*, *CACNA1A*, *CDH18*, *CDH9*, *CSMD1*, *CTNND2*, *FKBP9L*, *GRM3*, *HCN1*, *KDM5A*, *LANCL2*, *MYO10*, *NNT*, *PARP8*, *PCDH11X*, *RNA5SP251*, *SEC61G*, *SMARCA4*, *TPTE*, *UBR4*, *VOPP1*, *VSTM2A*, *ZNF236*, *ZNF713*
40	*AGTR1*, *ATR*, *BCHE*, *C3orf33*, *C3orf55*, *C3orf79*, *CHST2*, *CLRN1*, *CLSTN2*, *CP*, *CPA3*, *CPB1*, *CT64*, *DHX36*, *EIF2A*, *FAM188B2*, *FAM194A*, *GFM1*, *GOLIM4*, *GPR149*, *GPR87*, *GRK7*, *GYG1*, *HLTF*, *HPS3*, *IGSF10*, *IL12A*, *KALRN*, *KCNAB1*, *KPNA4*, *LINC00886*, *MFSD1*, *MIR1263*, *MIR15B*, *MIR720*, *MLF1*, *NMD3*, *OTOL1*, *P2RY1*, *P2RY12*, *P2RY13*, *PFN2*, *PIK3CB*, *PLCH1*, *PLOD2*, *PLSCR5*, *RN7SKP298*, *RNF13*, *RSRC1*, *SCHIP1*, *SELT*, *SERP1*, *SHOX2*, *SI*, *SIAH2*, *SLC33A1*, *SLC9A9*, *SLITRK3*, *SMC4*, *SPTSSB*, *SSR3*, *STAG1*, *SUCNR1*, *TIPARP*, *TM4SF1*, *TM4SF4*, *TRIM59*, *TRPC1*, *TSC22D2*, *VEPH1*, *WDR49*, *WWTR1*, *ZBBX*, *ZIC1*, *ZIC4*

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:0002323	Natural killer cell activation involved in immune response	6.02e-29
	GO:0042100	B cell proliferation	1.65e-26
	GO:0043330	Response to exogenous dsrna	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:0010469	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:0000904	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

Table S4. List of candidate drivers whose corresponding SGAs are 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	*NKRF*
3	***RB1***	28	***AHNAK***	53	*EP400*	78	***SETD2***
4	***PBRM1***	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	*ANKRD30BP2*	82	*WRN*
8	***KRAS***	33	*OCA2*	58	*PRICKLE2*	83	*LHX1*
9	*SLC1A6*	34	***GATA3***	59	*RGAG1*	84	***STAG2***
10	*POMC*	35	*PCNX*	60	*KRT23*	85	*KSR1*
11	*RRN3P2*	36	*KRTAP4-9*	61	*UGT1A1*	86	*GCDH*
12	*TFAM*	37	*LRRIQ3*	62	*PARP8*	87	*E2F3*
13	*CD163*	38	*MRGPRF*	63	*TMPRSS6*	88	*PDHX*
14	*WDFY3*	39	***HSP90AA1***	64	*FMN2*	89	*CLUH*
15	*WDR44*	40	*CNTN3*	65	***CDKN2A***	90	*PRICKLE4*
16	*CYP51A1*	41	*WNK3*	66	*DIP2B*	91	*GLUD2*
17	*ADARB2*	42	*PTPRD*	67	*TBP*	92	*CROCC*
18	*C9orf53*	43	*PCDHB16*	68	*ZNF624*	93	***IDH1***
19	***BAP1***	44	*RPLP0P2*	69	*FEM1B*	94	*GRIA1*
20	*TMPRSS13*	45	*COL6A1*	70	*CDKN2B*	95	*DLG5*
21	*SV2C*	46	*TTC39B*	71	*PDE4D*	96	*SMURF2P1*
22	*MYCBP2*	47	***PGR***	72	*ISLR2*	97	*CACNA1C*
23	***MED24***	48	*TBC1D4*	73	*FLRT3*	98	*KIAA1377*
24	***CYLD***	49	*ANKRD36C*	74	*ZFAT*	99	*PTPRZ1*
25	*CYLC2*	50	*GPATCH8*	75	***SMARCA4***	100	***PCSK5***

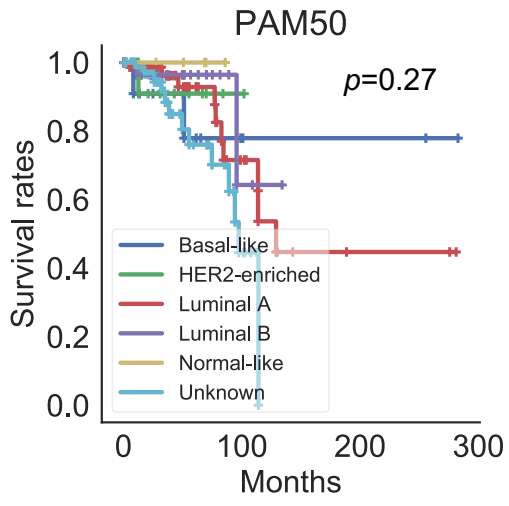

Fig. S7. Survival profiles of breast cancer subtypes inferred from PAM50. PAM50 is a widely used baseline for identifying subgroups of breast cancer, which classifies the breast cancers based on the expression profiles of the 50 most important genes that indicative of basal and myoepithelial features 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 5**d**, this reflects the utility of tumor embedding to represent tumor status.

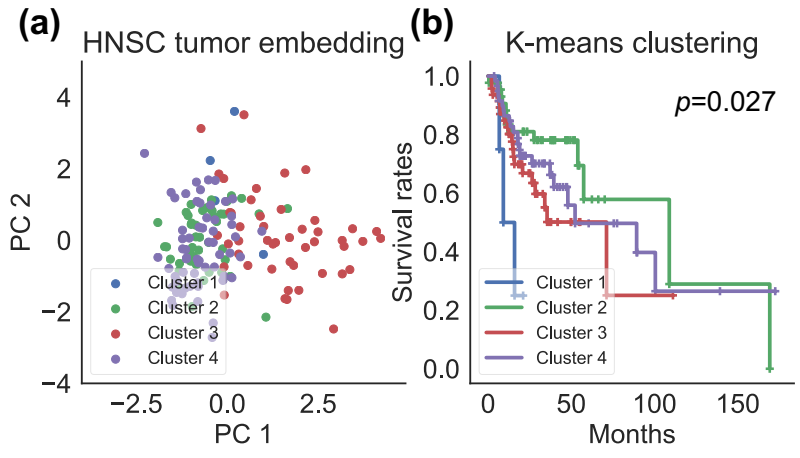

Fig. S8. Log-rank test on *k*-means clustering results of HNSC tumors represented by tumor embeddings. Additional evaluation implemented on the second largest corpus of cancer type (head neck squamous cell carcinoma; HNSC) in our dataset. It further validates the feasibility of tumor embedding in representing tumor status, in addition to the results of Figure 5**a,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).