OmicsMapNet: Transforming omics data to take advantage of Deep Convolutional Neural Network for discovery

Shiyong Ma, Zhen Zhang
Department of Pathology, Johns Hopkins Medical Institutions 1550 Orleans Street, Baltimore, MD 21231, USA

Abstract:

We developed OmicsMapNet approach to take advantage of existing deep leaning frameworks to analyze high-dimensional omics data as 2-dimensional images. The omics data of individual samples were first rearranged into 2D images in which molecular features related in functions, ontologies, or other relationships were organized in spatially adjacent and patterned locations. Deep learning neural networks were trained to classify the images. Molecular features informative of classes of different phenotypes were subsequently identified. As an example, we used the KEGG BRITE database to rearrange RNA-Seq expression data of TCGA diffuse glioma samples as treemaps to capture the functional hierarchical structure of genes in 2D images. Deep Convolutional Neural Networks (CNN) were derived using tools from TensorFlow to learn the grade of TCGA LGG and GBM samples with relatively high accuracy. The most contributory features in the trained CNN were confirmed in pathway analysis for their plausible functional involvement.

Keywords: Treemap, Omics Expression data, Deep Convolutional Neural Network, KEGG BRITE.

1. Introduction:

In the recent decades, high throughput proteogenomic analysis technologies have become much more mature (Nesvizhskii 2014). This has led to an explosion of genomics, transcriptomics, proteomics and other datasets from large-scale profiling efforts of biological/clinical samples. The analysis of genes and the products of genes contributes to deeper understandings of human diseases at molecular level. Molecular biological omics data are now serving as important resources in biomarkers research for initial discovery, further distillation, and possible collaborative verification (Ludwig and Weinstein 2005; Skates et al. 2013; Zhang, Tian, and Zhang 2014).

Currently, much of the omics data analysis for biomarker discovery involves some form of global univariate significance tests with adjustment/control for multiple-testing. It is often only until the number of potential targets has been significantly reduced that more complex multivariate approaches, such as GSEA (Mootha et al. 2003; Subramanian et al. 2005), differential dependency network analysis (Zhang et al. 2009), CAMERA (Wu and Smyth 2012) and GSVA (Hänzelmann, Castelo, and Guinney 2013), are used to assess their performance in combination. It is desirable therefore to develop supervised multivariate approaches that are capable of analyzing high-dimensional data of which features are sparsely related based on a prior knowledge and an undetermined subset of them could be potentially complementary in classification of samples in data.

Computationally, it is not straightforward to organize such data to allow for effective and efficient application of conventional multivariate analysis algorithms on omics data directly to select candidate biomarkers. We developed a new approach to rearrange high-dimensional omics data as 2-dimensional
images in which molecular features related in functions, ontologies, or other relationships were organized in spatially adjacent and patterned locations. We then take advantage of existing deep leaning frameworks to train classifier deep learning neural networks to classify the images based on known sample classes. The selection of potential biomarkers involved the identification of informative features for the trained neural network.

Specifically in this research, as an example, to rearrange the omics data and convert the data into 2-dimensional images, functional annotation of genes was employed and treemap approach was used. First, KEGG (http://www.kegg.jp/) is a database of interpreting genome sequence and other high-throughput data biologically (Minoru Kanehisa et al. 2016), it aims at linking sets of genes in the genomes to biological functions of the cell and the organism. Hierarchical tree structure is a way of presenting the complex biological functions of molecules, KEGG BRITE is reference database which contains a set of manually created hierarchical texts. And those texts capture the functional hierarchies of biological objects, especially the KEGG objects (M. Kanehisa 2006; Minoru Kanehisa et al. 2012, 2008). So KEGG BRITE hierarchical structure was used for the hierarchical functional annotations. Second, treemap is a space-filling method for visualizing large scale hierarchical structures (Shneiderman 1992; Bederson, Shneiderman, and Wattenberg 2002), and was used for converting the generated hierarchical structures into 2-dimensional images. Some other methods of potentially converting omics expression data into 2-dimensional images have also been proposed previously, and those methods include using heatmap to visualize bi-clustering (Barkow et al. 2006) and covariant matrix (González et al. 2012).

Previously, treemap methods for visualizing omics expression data were proposed. For example, Baehrecke applied treemaps to visualize the microarray based gene expression data in the context of GO hierarchical annotations (Baehrecke et al. 2004), Halligan used treemaps to display quantitative proteomics data using simplified version of GO annotation - Generic GO slim (Halligan et al. 2007), and Bernhardt (Bernhardt et al. 2009) visualized gene expression data via Voronoi treemaps using GO slim generic and KEGG Orthology. Yet, compared to Halligan and Baehrecke’s work, this current work removed the labels of the units and uninformative boundaries, and each gene in the treemaps was assigned with the same size of area. Compared to Bernhardt’s work, the creation procedures of treemaps are similar, yet we used rectangle treemaps instead of Voronoi treemaps. In summary, we customized the process and aim at creating treemaps that are suitable for deep learning model to learn.

In this current work, as an example, we explored a TCGA dataset to create treemap images. We obtained the gene expression matrix of TCGA diffuse glioma samples from Ceccarelli’s paper (Ceccarelli et al. 2016) at https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/. The grading of gliomas is important as erroneous grading may lead to either over- or under-treatment, particular affects the radiotherapy dosage and the use of chemotherapy. The grading of gliomas from WHO were traditionally classified based on histologic type and malignancy grade (Louis et al. 2007), and in 2016 the grading from WHO further takes several molecular features (IDH-mutant, 1p/19q-codeleted, etc.) into consideration to define several new gliomas.

Convolutional Neural Networks was inspired by visual cortex of the brain, and are designed to process 2-dimensional images (Krizhevsky, Sutskever, and Geoffre E. 2012). Therefore, we designed deep Convolutional Neural Network (CNN) to learn and understand the created treemap images. In this work, the CNN was written in python employing TensorFlow framework. The developed CNN was trained to learn and predict the grade of the TCGA LGG and GBM samples, the contributory genes for the classification were obtained from the CNN feature maps. Pathway analysis was further done on the selected contributory genes from the feature maps, and the analysis results show that plausible functional pathways are enriched.
2. Methods:

2.1 Omics expression data

In this work, the RNA-Seq expression matrix of the TCGA LGG&GBM dataset was obtained at [https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/LGG-GBM.gene_expression.normalized.txt](https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015/LGG-GBM.gene_expression.normalized.txt). To be specific, we first used the R package edgeR (Robinson, McCarthy, and Smyth 2010) to normalize the RAW count number into TMM value (log2) (Risso et al. 2014). Then we filtered out the genes with extremely low TMM values, since it is assumed that extremely low abundant genes are likely not to influence the biological processes. The threshold of TMM value that we used for filtration is -5. These gene names were matched to KEGG IDs using R package UniProt.ws from Bioconductor.

2.2 Transforming Omics expression data into 2-dimensional images

In this work, to transform omics expression data into 2-dimensional images, we first parsed the KEGG BRITE structure at [http://rest.kegg.jp/get/br:br08902](http://rest.kegg.jp/get/br:br08902). In this work, we only considered the Genes and Proteins branch of the BRITE hierarchical structure. A two-layer hierarchical tree with 58 child nodes was initially generated, and each child node represents a ko hierarchy file. Then we explored the ko hierarchy files and extracted the next layer of functional groupings from those files and updated the generated tree with a third layer, the finer tree contains 234 child nodes. Then we further assigned the genes to the corresponding child nodes of the tree structure, and finally generated a hierarchical tree with four layers. Considering that one gene might have multiple KEGG functional annotation, one gene might appear in several places in the tree.

Second, to spatially arrange the genes of a sample into a 2-dimensional image, rectangle treemaps were used. Rectangle units were used to represent genes in the treemaps. Considering that each rectangle unit represents one gene, each rectangle unit was assigned the same size of area. To put the rectangle units into the treemap, several algorithms have been proposed previously. In this paper, the pivot method (Bederson, Shneiderman, and Wattenberg 2002) was used. To realize this, the R package TreeMap from CRAN was obtained and modified. Furthermore, within each functional group, the genes were sorted according to the median value of the TMM (log2) values across all the samples. All the treemap images have exactly the same spatial arrangement of the genes.

Third, we colored the treemaps based on the normalized value of the gene expression abundances. The normalized gene expression values were mapped to colors in the BlueYellowRed heatmap sample by sample. For each sample, the highest TMM value (log2) was mapped to the Red color, and the lowest TMM value (log2) was mapped to the Blue color, 256 colors were interpolated, TMM values (log2) were linearly mapped to the corresponding colors in the BlueYellowRed heatmap. An advantage of this method is that the relative abundance of different genes within one sample are kept.

The created treemap images contains 1024*1024 pixels, the images were subsampled to 512*512 pixels before deep CNN. An example of created treemap image was given in Figure 1 layer by layer.

2.3 The designed deep Convolutional Neural Network

The architecture of the designed deep CNN is shown in Figure 2. It is with 3 convolutional layers (conv), 2 fully connected layers (fc), the RELU (Tang 2013) activation functions was used, Adam Optimizer (Kingma and Ba 2014) was used as the optimizer, and the loss function is Cross Entropy. Conventional convolution neural network layer normally contains alternating CNN layers and pooling layers, this
method was adopted in this work. Max Pooling method was used as the pooling function. For the conv layers, the filter size is 3, the first and second conv layer has 32 filters respectively, and the third conv layer has 64 filters, the strides of all the conv filters are 1, and the strides for all the pooling layers are 2, the padding method adopted is “VALID”. The first FC layer and second FC layer have 128 and 3 neurons respectively. For other hyper-parameters of the model, dropout method which was proposed by Hinton (Srivastava et al. 2014) was used on FC1 layer (dropout rate is 0.75), L2 regularization was used for both FC layers to further reduce overfitting and the beta parameter is 0.01, learning rate is 0.001, batch size for learning is 25, the number of iterations is 500 and early stopping number is 10. 20-fold cross validation method was adopted, treemap images of all the sample were randomly divided into 20 folds (fold size = 30 or 31).

### 2.4 Selection of contributory gene set based on the feature maps of CNN

Apart from the overall prediction accuracy of the model, we are further interested in using the deep CNN model to select or assume a set of genes that are highly contributory to the classification. We noticed that the sum of pixel intensity of feature maps reduces from low conv layer to high conv layer. In conv3, most of the features maps rarely have any nonzero pixel intensities. Even for the several strongest feature maps (the maps whose sum of pixel intensities are the largest), they are likely to be sparse. In this work, since conv3 layer is directly linked to the downstream fully connected layer, the feature maps generated from conv3 layer were used for this gene set selection. To be specific, conv3 layer has 64 channels, so 64 feature maps were outputted from conv3, and we selected the strongest feature map (which has the largest sum of pixel intensities).

An example of the selected feature map is shown in Figure 3. Since signal intensity of conv3 feature map is low, heatmaps were used to visualize the feature maps in this paper. In computer vision, for object detections, by projecting the strong responses in feature maps back to the original image (He et al. 2015), reasonable features could be potentially discovered. In this work, we adopted the same approach, and projected the strong responses to original treemap. Due to the pooling operation, the feature map is down sampled from original treemap. The size of feature map from conv3 is 62*62, so each pixel in the conv3 feature map was linearly projected to a square area with PR*PR pixels in the treemap image (PR = (1024/62)), the projection is shown in Figure 4. In this work, if the center of a gene’s rectangle is within the square area of selected response point, it is assumed that the selected response point can be projected to the gene.

To summarize the strong response points of the selected feature maps, the top 10% strongest points in each feature maps were selected. Then we calculated the number of selections of each pixels in the feature maps of all samples. To visualize the frequency of selection of each pixel, a heatmap was used as well. The genes which were able to stably providing strong response in the feature maps were obtained.

### 2.5 Pathway analyses of selected gene set

To further interpret the selected contributory genes, Cytoscape (Shannon et al. 2003) ClueGO (Bindea et al. 2009) was used. Pathway analysis was done on the selected genes. In this work, Gene Ontology Biological Process (Ashburner et al. 2000) and REACTOME pathway (Fabregat et al. 2016) analyses were done. For both pathways analyses, we chose the specific terms for analysis, used the GO term fusion for GO enrichment analysis, and pathways only with p-value smaller than 0.05 were shown in the figure.
3. Results:

3.1 The acquisition of gene data for treemap creation

From the TCGA LGG&GBM dataset, 20330 genes were initially obtained from the gene expression matrix, and 17715 genes were obtained after eliminating genes with extremely low abundances. 10896 genes of the 17715 filtered genes are able to be mapped to the KEGG IDs, therefore, these 10896 genes were used to generate the treemaps. Also, in the dataset, RNA-Seq analysis was done on 667 samples, and 607 samples of them were labelled with WHO grade (Louis et al. 2016, 2007). Therefore, in total, 607 treemap images were created.

3.2 The performance of predicting the grade using designed CNN and treemap image

For the 607 TCGA LGG&GBM samples, the number of samples of different grade are listed below:
- Grade-II: 215
- Grade-III: 239
- Grade-IV: 153

The performance of the CNN was assessed using 20-fold cross validation. Shapiro-Wilk test was done on the prediction accuracies of 20 folds, the result indicates that the accuracies follows a normal distribution (p-value = 0.9265). The mean value of the accuracies is 0.71543, the 95% confidence interval is from 0.66966 to 0.761193. A label permutation test was further done for comparison. Shapiro-Wilk test also indicates that the accuracies follows a normal distribution (p-value = 0.9782), the mean value of the prediction accuracies is 0.4025806, the 95% confidence interval is from 0.3660146 to 0.4391467. A Welch Two Sample t-test was done on prediction accuracies of labeled and label permutated samples, p-value is 2.676e-13, and the result indicates that the means are significantly different.

The ROC curves are shown in Figure 5. In the figure, G2 and G3 are relatively difficult to be distinguished, since the AUC of G2 is 0.82 and the AUC of G3 is 0.77. The AUC of G4 is 0.99, which indicates that G4 can be distinguished from G2 and G3 with high accuracy.

3.3 Analysis of the selected gene sets.

To investigate the properties of feature maps, we first calculated the sum of pixel intensities of the 64 conv3 feature maps of all the treemap images. The information was saved in Supplementary File 1 “1 ActivationMap_signalIntensitySummaryCNV3.csv”. For most samples, the pixel intensities were concentrated to several feature maps. From the heatmaps of the selected feature maps, we notice that the strong response points are sparse and therefore can only be projected to a subset of genes.

For the 607 samples, the numbers of gene selection in the feature maps were calculated and visualized in Figure 6. The strong response points are concentrated in certain areas of the treemap images. Using the projection method in 2.4, genes that correspond to the strong responses were obtained. The corresponding genes and number of selection of those genes are listed in Supplementary File 2 “2 A_SummaryofGenes_SelectRatio.xlsx”. From this file, genes with KEGG id “K02923”, “k02957” and ko03011 “Ribosome: Ribosomal proteins” functional annotation were selected in 428 samples.

We further explored genes relevant to LGG and GBM from literatures (Ceccarelli et al. 2016; Zheng et al. 2008; Eckel-Passow et al. 2015) the genes and corresponding KEGG id are listed in Table 1. The positions of these genes in the treemap image is given in Figure 7. We further check the numbers of
selections of these genes, yet, it shows that the selection times vary largely. For example, K00031 was selected 331 times, yet, K20225 was only selected for 25 times. And some genes were not selected at all.

One gene might appear in multiple places in the treemap image, this is because genes might have multiple KEGG functional annotations. Based on the local connectivity property of deep CNN model, so one gene might be selected in one functional position, the same gene in the other functional positions might not be selected. In brief, genes are selected locally based on functional groupings. For example, for the gene K05021, the pixel point with ko04040 “Ion Channels: Chloride channels” annotation term was selected 371 times, comparatively, the pixel point with ko04147 “Exosome: Exosomal proteins” annotation term was selected only 71 times.

3.4 Pathway analyses of selected gene set.

We further analyzed the pathways of the highly frequently selected genes. The genes with at least 300 selection number were selected. In total, 485 genes were selected, the gene IDs are listed in Supplementary File 3 “3 ASelected_GeneListConverted.csv”. In Figure 8, Gene Ontology (Biological Process) pathway enrichment analysis result is illustrated. Oculomotor nerve formation pathway (p = 3.6e-3) which is related to cranial nerve formation pathway are enriched. In Figure 9, REACTOME pathways enrichment analysis is illustrated, pathways like Metabolism of serotonin (p = 1.1e-3) and Loss of Function of TGFBR2 in cancer (p = 6.6e-3) are enriched. Several enriched functional annotations are associated with nerve systems and/or cancer progression.

4. Conclusion and discussion:

Omics expression dataset (e.g proteomics data, mRNA expression data) usually contains thousands of genes or proteins, the dimensions are much larger than the number of samples. To discover effective biomarkers for certain phenotypes or clinical properties, strict feature selection from the high dimensional data is required. In recent years, deep learning is proved to have strong capacity in discovering intricate patterns in high-dimensional data (Lecun, Bengio, and Hinton 2015), with this consideration, we developed OmicsMapNet approach. The aim is to convert the high-dimensional omics data into a format on which deep learning can be used for feature discovery.

Specifically, since deep CNN has been widely used for visual understanding successfully (Ranzato et al. 2006; Szegedy et al. 2015; Simonyan and Zisserman 2015), in this current work, we transformed omics expression data into 2-dimensional images and developed the deep CNN model to understand the grade of TCGA LGG and GBM samples. Traditional approaches rely heavily on global univariate analysis, yet, the deep CNN pooled the original treemap images three times. During spatial pooling method, the outputs of several nearby feature detectors are combined into local or global “bag of features”, the aim is to enable compact representations, better robustness to noise (Boureau, Ponce, and LeCun 2010). Therefore, in this work, the deep CNN provides a way to jointly considering multiple variant locally with similar functions.

To improve OmicsMapNet approach in this work, there are several issues that need further investigating and studying. For example, in these selected feature maps, the genes with non-zero pixel intensities are more densely located in the “cold” areas in the heatmap. It indicates that some biases exist. Essential normalization methods could be developed to remove the biases. Moreover, considering that gene semantic annotations are still facing multiple challenges, updating the functional groupings or using finer functional annotations might help to improve the prediction performance and selecting of gene set.
Finally, although the RNA-Seq expression data was used to generate the treemap images in this work, yet, OmicsMapNet method can also be used for protein expression data and other types of quantitative omics and clinical data.

**List of Abbreviations:**

- ATRX - ATP-dependent X-linked helicase
- CIC - Drosophila homolog of capicua
- EGFR - Epidermal growth factor receptor
- GB IV - Glioblastoma WHO grade IV
- GO - Gene ontology
- IDH - Isocitrate dehydrogenase
- PTEN - Phosphatase and tensin homolog on chromosome 10
- TERT - Telomerase reverse transcriptase
- TP53 - Tumor protein p53
- WHO - World Health Organization

**Code Availability:**

Please contact authors.

**References:**

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics*. https://doi.org/10.1038/75556.

Bachrecke, Eric H., Niem Dang, Ketan Babaria, and Ben Shneiderman. 2004. “Visualization and Analysis of Microarray and Gene Ontology Data with Treemaps.” *BMC Bioinformatics* 5. https://doi.org/10.1186/1471-2105-5-84.

Barkow, Simon, Stefan Bleuler, Amelia Prelić, Philip Zimmermann, and Eckart Zitzler. 2006. “BiC: A Biclustering Analysis Toolbox.” *Bioinformatics* 22 (10):1282–83. https://doi.org/10.1093/bioinformatics/btl099.

Bederson, Benjamin B., Ben Shneiderman, and Martin Wattenberg. 2002. “Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies.” *ACM Transactions on Graphics* 21 (4):833–54. https://doi.org/10.1145/571647.571649.

Bernhardt, Jörg, Stefan Funke, Michael Hecker, and Juliane Siebourg. 2009. “Visualizing Gene Expression Data via Voronoi Treemaps.” In *6th International Symposium on Voronoi Diagrams in Science and Engineering, ISVD 2009*, 233–41. https://doi.org/10.1109/ISVD.2009.33.

Bindea, Gabriela, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. 2009. “ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks.” *Bioinformatics* 25 (8):1091–93. https://doi.org/10.1093/bioinformatics/btp101.

Boureau, Y-Lan, Jean Ponce, and Yann LeCun. 2010. “A Theoretical Analysis of Feature Pooling in Visual Recognition.” *Icmil*, 111–18. https://doi.org/citeulike-article-id:8496352.

Ceccarelli, Michele, Floris P. Barthel, Tathiane M. Malta, Thais S. Sabetot, Sofie R. Salama, Bradley A. Murray, Olena Morozova, et al. 2016. “Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.” *Cell* 164 (3):550–63.
Eckel-Passow, J E, D H Lachance, A M Molinaro, K M Walsh, P A Decker, H Sicotte, M Pekmezci, et al. 2015. “Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors.” New England Journal of Medicine 372 (26):2499–2508. https://doi.org/10.1056/NEJMoA1407279.

Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. “The Reactome Pathway Knowledgebase.” Nucleic Acids Research 44 (D1):D481–87. https://doi.org/10.1093/nar/gkv1351.

González, Ignacio, Kim Anh Lê Cao, Melissa J. Davis, and Sébastien Déjean. 2012. “Visualising Associations between Paired ‘Omics’ Data Sets.” BioData Mining 5 (1). https://doi.org/10.1186/1756-0381-5-19.

Halligan, Brian D., Shama P. Mirza, Molly C. Pellitteri-Hahn, Michael Olivier, and Andrew S. Greene. 2007. “Visualizing Quantitative Proteomics Datasets Using Treemaps.” In Proceedings of the International Conference on Information Visualisation, 527–34. https://doi.org/10.1109/IV.2007.136.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. “GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data.” BMC Bioinformatics 14. https://doi.org/10.1186/1471-2105-14-7.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.” IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9):1904–16. https://doi.org/10.1109/TPAMI.2015.2389824.

Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, et al. 2008. “KEGG for Linking Genomes to Life and the Environment.” Nucleic Acids Research 36 (SUPPL. 1). https://doi.org/10.1093/nar/gkm882.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2016. “KEGG as a Reference Resource for Gene and Protein Annotation.” Nucleic Acids Research 44 (D1):D457–62. https://doi.org/10.1093/nar/gkv1070.

Kingma, Diederik, and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” arXiv Preprint arXiv:1412.6980, 1–13. https://doi.org/10.1109/ICCE.2017.7889386.

Krizhevsky, Alex, Ilya Sutskever, and Hinton Geoffrey E. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” Advances in Neural Information Processing Systems 25 (NIPS2012), 1–9. https://doi.org/10.1109/5.726791.

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” Nature. https://doi.org/10.1038/nature14539.

Louis, David N., Hiroko Ohgaki, Otmar D. Wiestler, Webster K. Cavenee, Peter C. Burger, Anne Jouvet, Bernd W. Scheithauer, and Paul Kleihues. 2007. “The 2007 WHO Classification of Tumours of the Central Nervous System.” Acta Neuropathologica. https://doi.org/10.1007/s00401-007-0243-4.

Louis, David N., Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. 2016. “The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary.” Acta Neuropathologica. https://doi.org/10.1007/s00401-016-1545-1.

Ludwig, Joseph A., and John N. Weinstein. 2005. “Biomarkers in Cancer Staging, Prognosis and Treatment Selection.” Nature Reviews Cancer. https://doi.org/10.1038/nrc1739.

Mootha, Vamsi K., Cecilia M. Lindgren, Karl Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. “PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinate Downregulated in Human Diabetes.” Nature Genetics 34 (3):267–73. https://doi.org/10.1038/ng1180.

Nesvizhskii, Alexey I. 2014. “Proteogenomics: Concepts, Applications and Computational Strategies.”
Ranzato, Marc Aurelio, Fu Jie Huang, Y-lan Boureau, Yann Lecun, Christopher Poultney, Sumit Chopra, and Yann L. Cun. 2006. “Efficient Learning of Sparse Representations with an Energy-Based Model.” *Advances in Neural Information Processing Systems* 19:1137–44.

Risso, Davide, John Ngai, Terence P. Speed, and Sandrine Dudoit. 2014. “Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples.” *Nature Biotechnology* 32 (9):896–902. https://doi.org/10.1038/nbt.2931.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics (Oxford, England)* 26 (1):139–40. https://doi.org/10.1093/bioinformatics/btp616.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11):2498–2504. https://doi.org/10.1101/gr.1239303.

Skates, Steven J., Michael A. Gillette, Joshua LaBaer, Steven A. Carr, Leigh Anderson, Daniel C. Liebler, David Ransohoff, et al. 2013. “Statistical Design for Biospecimen Cohort Size in Proteomics-Based Biomarker Discovery and Verification Studies.” *Journal of Proteome Research* 12 (12):5383–94. https://doi.org/10.1021/pr400132j.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” *Journal of Machine Learning Research* 15:1929–58. https://doi.org/10.1214/12-AOS1000.

Subramanian, A, P Tamayo, V K Mootha, S Mukherjee, B L Ebert, M A Gillette, A Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proc Natl Acad Sci USA* 102 (43):15545–50. https://doi.org/10.1073/pnas.0506580102.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going Deeper with Convolutions.” In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07–12–June:1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Tang, Yichuan. 2013. “Deep Learning Using Linear Support Vector Machines.” *arXiv:1306.0239 [Cs, Stat]*. http://arxiv.org/abs/1306.0239.

Wu, Di, and Gordon K. Smyth. 2012. “Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation.” *Nucleic Acids Research* 40 (17). https://doi.org/10.1093/nar/gks461.

Zhang, Bai, Huai Li, Rebecca B. Riggins, Ming Zhan, Jianhua Xuan, Zhen Zhang, Eric P. Hoffman, Robert Clarke, and Yue Wang. 2009. “Differential Dependency Network Analysis to Identify Condition-Specific Topological Changes in Biological Networks.” *Bioinformatics* 25 (4):526–32. https://doi.org/10.1093/bioinformatics/btn660.

Zhang, Bai, Ye Tian, and Zhen Zhang. 2014. “Network Biology in Medicine and beyond.” *Circulation: Cardiovascular Genetics* 7 (4):536–47. https://doi.org/10.1161/CIRCGENETICS.113.00123.

Zheng, Hongwu, Haoqiang Ying, Haiyan Yan, Alec C. Kimmelman, David J. Hiller, An Jou Chen, Samuel R. Perry, et al. 2008. “p53 and Pten Control Neural and Glioma Stem/progenitor Cell Renewal and Differentiation.” *Nature* 455 (7216):1129–33. https://doi.org/10.1038/nature07443.
Figures:

Figure 1(a)
Figure 1(b)
Figure 1(d)
An example of treemap image. The hierarchical structure is shown layer by layer. Each gene rectangle represent one gene, so they have the same size of area. Therefore, the size of the rectangle of functional unit represents the number of genes within the unit. The colors of the gene rectangle represent the normalized gene abundances, yet the colors of the functional unit rectangles were randomly generated.
Figure 2. The architecture of deep CNN

Figure 3. An example of the Feature Map of Conv3
Figure 4. Projecting pixel in feature map to treemap image.

Figure 5. ROC curves of the deep CNN
Figure 6. Selection the strongest response pixels from the feature maps across all the samples, and plot the selection frequency using a heatmap.

Figure 7. Positions of Key Genes. The Red Rectangles represents the key genes inferred from literatures.
Figure 8. GO (Biological Process) analysis of selected genes
Figure 9. REACTOME pathway analysis of selected genes

Tables:

Table 1. Important Genes according to relevant literatures: Gene Name and corresponding KEGG entry ID:
ARID2 - K11765
ATRX – K10779
CIC - K20225
DNMT3A - K17398
EGFR - K04361
FUBP1 - K13210
IDH1 - K00031
IDH2 - K00031
IDH3A - K00030
IDH3B - K00030
IDH3G - K00030
KRAS - K07827
NRAS - K07828
PTEN – K01110
SETD2 - K11423
TERT – K11126
TP53 - K04451