MOLECULAR BASES OF MORPHOMETRIC COMPOSITION IN GLIOBLASTOMA MULTIFORME

Ju Han[1]*, Hang Chang*[1]*, Gerald V. Fontenay*[1]*, Paul T. Spellman*[2]*, Alexander Borowsky*[3]*, and Bahram Parvin*[1]

[1] Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.
[2] Center for Spatial Systems Biomedicine, Oregon Health Sciences University, Portland, Oregon, U.S.A.
[3] Center for Comparative Medicine, University of California, Davis, California, U.S.A.

ABSTRACT

Integrated analysis of tissue histology with the genome-wide array (e.g., OMIC) and clinical data have the potential for hypothesis generation and be prognostic. OMIC and clinical data are typically characterized and summarized at the patient level while whole mount histological sections are often heterogeneous in terms of nuclear morphology and organization. In this paper, we propose a multi-level framework for summarization and association of morphometric data. At the lowest level, each nucleus is segmented and then profiled with a multi-dimensional representation. At the intermediate level, cellular profiles are summarized within a local neighborhood, and further clustered into subtypes. At the highest level, each patient is represented by the composition of subtypes that are computed from the intermediate level, and then integrated with OMIC and outcome data for further analysis. The framework has been applied to Glioblastoma multiforme (GBM) data from The Cancer Genome Atlas (TCGA). Based on cellularity and nuclear size, four subtypes have been identified at the intermediate level. Subsequent multi-variate survival analysis indicates that the patient composition of one of the subtypes, with extremely low cellularity and small nucleus size, has a significantly higher hazard ratio. Further correlation of this subtype with the molecular data reveals enrichment of (i) STAT3 pathway and (ii) common regulators of PKC, TNF, AGT, and PDGF.

***Index Terms*—** Tumor architecture, Cox proportional-hazards model, consensus clustering, molecular association

1. INTRODUCTION

The Cancer Genome Atlas (TCGA) is a national collaborative effort that aims to identify molecular aberrations of tumors. This enterprise is also coupled with a collection of histology sections from regions adjacent to the biopsies for diagnostics. While genome-wide molecular data have the advantages of being structured, histological sections are often unstructured and difficult to process. However, histological sections provide rich phenotypic information, such as tumor architecture and heterogeneity. Various techniques for utilizing histological sections for tumor grading and association with the clinical outcomes was briefly summarized in [1].

Recent efforts on TCGA histological data analysis have focused on normalizing for the batch effect [2], developing a computational pipeline to process whole mount histology sections [1, 3], and associating tumor histology with molecular data [1, 4]. This paper focuses on a new perspective for identifying molecular drivers of tumor composition that can only be realized when processing a large cohort of whole-mount tissue sections. In our processing pipeline, each whole mount tissue section is decomposed into blocks of 1k×1k pixels. To a first approximation, these blocks serve as the bases for intermediate-level analysis and are represented by probability distributions of morphometric indices aggregated from the cell-by-cell analysis. Computed subtypes from these probability distributions at the intermediate level serve as a "coding" procedure. Subsequently, at the patient level, the tumor architecture is represented by the composition of subtypes that are computed from the intermediate level, and then integrated with OMIC and clinical data for further analysis.

Organization of the rest of this paper are as follows: Section 2 describes the technical details of our approach; Section 3 provides experimental results on integrated analysis of TCGA Glioblastoma multiforme data; and Section 4 concludes the paper.

2. APPROACH

The proposed hierarchical approach includes analysis at three levels: segmentation and feature extraction at the cellular level, feature summarization and subtyping at the intermediate level, and survival analysis and genomic association at the patient level.

2.1. Batch invariant analysis at the cellular level

One of the major challenges in histological image analysis is that tissue sections originate from different laboratories and are subject to a significant amount of technical variations. We have developed a novel approach for nuclear segmentation in tissue sections, which addresses the problem of technical and biological variations by incorporating information from manually annotated reference images [2]. Segmented nuclei enable a multi-dimensional representation that captures morphology and organization. These computed indices are then mined for subtyping at the intermediate level.

2.2. Morphometric summarization and subtyping at the intermediate level

For the purpose of intermediate level analysis, we divide each tissue section into non-overlapping regions of 1000×1000 pixels at the $20X$ resolution, each of which is called a block. Each feature is summarized as a probability distribution per block, and then it is normalized across all tissues within a tumor type.

This work was supported by NIH grant a U24 CA1437991 (ps) carried out at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231.

Morphometric subtyping of blocks across all tissue sections of a given tumor type enables subsequent survival and compositional analysis. However, several barriers need to be addressed for block-level subtyping: (i) blocks in the background or at the border of a tissue section may have undesired effects on subtyping; (ii) the number of blocks per whole mount tissue section is quite large (e.g., 2500); and (iii) whole mount tissue sections vary in size, as a result, leading to an imbalanced number of blocks per section.

To address these issues, we have developed a computational pipeline consisting of four major steps: 1) *filtering* each tissue section to remove background and border blocks, 2) *sampling* each tissue section to identify representative blocks, 3) *clustering* these representative blocks across all tissue sections of a given tumor type, and 4) *labeling* all remaining blocks based on clustering results of the representative blocks. The details of each step is summarized below.

1) In the *filtering* step, the background regions in a tissue section are first detected at a very low resolution. Any block containing the background region is then marked and removed from subsequent analysis.

2) *Sampling:* Initially, k-means algorithm is applied to identify morphometric clusters **within** each whole mount tissue section. The number of clusters k is selected proportional to the number of blocks in the tissue section (e.g., 1% of the number of blocks). Subsequently, the block, closest to the centroid of the cluster, is selected as a representative block for that cluster. In other words, k blocks are selected to represent each tissue section.

3) *Clustering:* Consensus clustering [5] is performed for identifying subtypes/clusters **across** tissue sections of a given tumor type. The input of consensus clustering includes blocks sampled from all tissue sections in the previous step. Consensus clustering aggregates consensus across multiple runs for a base clustering algorithm. Moreover, it provides a visualization tool to explore the number of clusters in the data, as well as assessing the stability of the discovered clusters. To remove blocks that are not appropriately clustered, we adapt a silhouette analysis method [6] that was used in [7]. The silhouette value for each block, normalized between -1 and $+1$, is a relative measure of how similar that block is to blocks in its own cluster compared to blocks in other clusters. A silhouette value close to $+1$ indicates that the block is appropriately clustered. Here, only blocks with positive silhouette values were retained as the training samples for the subsequent *labeling* step.

4) The final step is *labeling* (classifying), where each non-representative block is assigned to a cluster through nearest-neighbor classifier based on the training blocks.

2.3. Integrated analysis at the patient level

The results from the intermediate level subtyping enable a compositional representation at the patient level in which a patient has a certain percentage of blocks for each subtype. Tumor compositional covariates, at the patient level, can then be correlated with clinical covariates or genomic data for integrated analysis. For example, one of the questions aims to explore the relationship between the compositional covariates and survival distribution. In multivariate survival analysis, this relationship is typically examined through a parametric model [8]:

$$h(t) = \exp\left(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\right) \quad (1)$$

where $h(t)$ is the hazard function, the X's are the covariates, and the constant α represents a kind of log-baseline hazard. Without specifying the baseline hazard function $\alpha(t) = \log h0(t)$, the Cox proportional hazards (PH) model of

$$h(t) = h0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\right) \quad (2)$$

can be estimated by the partial likelihood method.

The Cox PH model was used to explore the relationship between the survival distribution and compositional covariates in the presence of important clinical covariates (e.g., age at initial pathologic diagnosis):

$$h(t) = h0(t) \exp(\beta_1 C_1 + \beta_2 C_2 + ...+ \beta_{N-1} C_{N-1} + \beta_N Age) \quad (3)$$

In Equation (3), C_i is the percentage of blocks belonging to the i-th subtype in all tissue sections from the same patient. Only $N-1$ compositional covariates are included in this model because the Nth covariate is linearly dependent on the other $N-1$ covariates. In the fitted model, covariates with small p-values are identified as statistically significant predictors of survival distribution.

With the identified histological covariate from the Cox PH model, we can now infer molecular candidates that best correlate with respect to the covariate. Pearson's product moment correlation coefficient was then computed between the histological covariate and expression values of each probeset for all available patients. The test statistic for assessing the significance of the correlation follows a t distribution with $n-2$ degrees of freedom, where n is the number of patients. P-values for the two tailed t-test were computed for all probesets, and then corrected for multiple testing using a false discovery rate (FDR) [9].

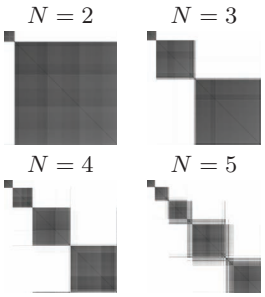


Fig. 1. Consensus clustering matrix of 146 TCGA patients with GBM for cluster number $N = 2$ to $N = 5$.

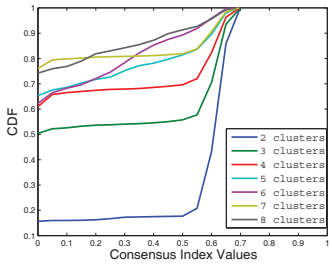


Fig. 2. Consensus clustering CDF for cluster number $N = 2$ to $N = 8$.

3. EXPERIMENTAL RESULTS

We have applied the above approach to computed representation of TCGA histology data of Glioblastoma multiforme (GBM), including 446 tissue sections from 152 patients. All the tissue sections

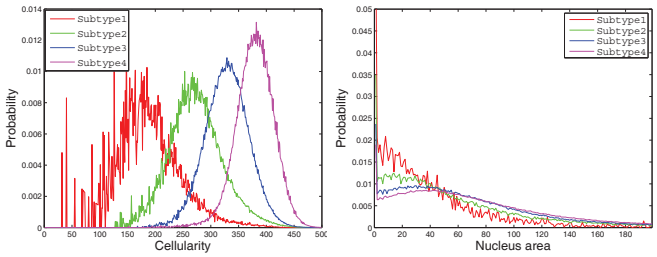


Fig. 3. Average equal-bin-width histograms of cellularity and nuclear size for each block-level subtype ($N = 4$).

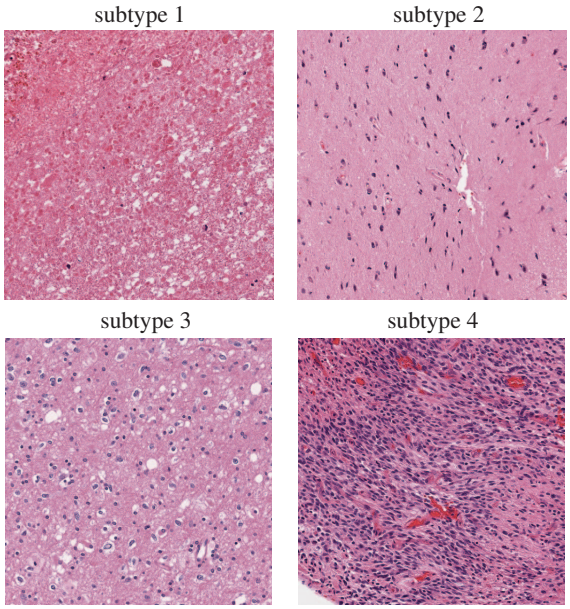


Fig. 4. Representative blocks for each morphometric subtype. Each block is of 1000-by-1000 pixels at 20X resolution.

were included for nuclear segmentation and morphometric representation. However, tissue sections that contain large blurred areas, penmarked areas, folds, and staining artifacts were removed from further analysis. The final dataset included 377 tissue sections from 146 patients. Since cellularity and nuclear size are important prognostic indices in GBM, we decided to explore them first for morphometic subtyping, survival analysis and genomic association. Gene expression values were estimated from the Affymetrix $HT_HG-U133A$ platform by Broad Institute.

3.1. Consensus clustering

Our representation for each block consists of a 25-bin equal probability histogram for nuclear size followed by a 25-bin equal probability histogram for cellularity. We identified $162,510$ non-background and non-border blocks through the filtering step, and selected $1,582$ representative blocks through the sampling step. We then use k-means algorithms as the base for consensus clustering, where the distance metric is the squared Euclidean distance. The procedure was run for 200 iterations with a sampling rate of 0.8 on $1,582$ representative blocks.

Consensus clustering matrices and CDFs (cumulative density functions), shown in Figure 1 and Figure 2, respectively, reveal four robust clusters (clustering stability significantly decreases for $N > 4$). We retained 1535 base blocks with positive silhouette as training samples for labeling all other blocks. Figure 3 shows the average equal-bin-width histograms of base blocks from each subtype. Four subtypes exhibit significantly different signatures in cellularity, i.e., subtypes 1 to 4 correspond to extremely low, low, mid and high cellularity, respectively. Similarly, subtypes 1 to 4 exhibits a monotonically increasing trend in nuclear size. Examples of representative blocks from each subtype are shown in Figure 4.

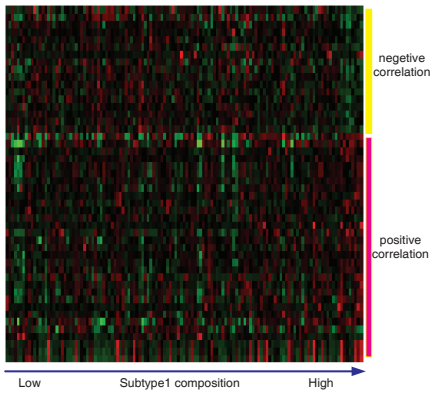


Fig. 5. Heatmap of top 48 probesets (rows) that best correlate with the subtype1 composition, with FDR adjusted p-value < 0.02.

3.2. Survival analysis and genomic association

Block labeling enables a compositional representation of each patient in which the percentage of blocks labeled as one of the subtypes. The relationship of the patient composition to the survival distribution is then examined through the Cox PH model in Equation (3). Tumor composition, in terms of computed subtypes and age, are then modeled as independent prognostic factors for patients.

Survival analysis is implemented through the R *survival* package. As mentioned in Section 2.3, we cannot incorporate all 4 compositional covariates into the Cox PH model simultaneously because they are linearly dependent. Instead, we choose 3 compositional covariates at a time for survival analysis. The results are summarized in Table 1. It is shown that both age and C1 (Subtype1 composition) have consistently high hazard ratio with p-values < 0.1 (i.e., these covariates are negatively correlated with survival). Block subtype1 has a histological signature of extremely low cellularity and small nuclear size, as shown in Figures 3 and 4, which is similar to the tumor signature of necrosis. This result is consistent with previous literature which shows that the extent of necrosis is negatively correlated with survival in GBM [10]. A heatmap of the 48 probesets that significantly correlate with the subtype1 composition, with FDR adjusted p-value < 0.02, are shown in Figure 5. These probesets were mapped into genes for further analysis.

Having identified genes that significantly correlate with the subtype1 composition, we then performed pathway and subnetwork enrichment analysis (see Figure 6). Pathway enrichment revealed STAT3, which is known to be a master regulator in GBM [11, 12]. Subnetwork enrichment identified AGT, PKC, PDGF, CEBPA, and TNF as the major hubs. Patients in this cohort received Temozolomide (TMZ) as a part of their treatment, which interferes with DNA replication through methylation. However, some tumor cells are able

Table 1. Multivariate survival analysis results by fitting the Cox PH model.

	Covariates in the Cox PH model							
	C1+C2+C3+Age		C1+C2+C4+Age		C1+C3+C4+Age		C2+C3+C4+Age	
	Hazard ratio	p-value	Hazard ratio	p-value	Hazard ratio	p-value	Hazard ratio	p-value
C1	1.0184	**0.0652**	1.0168	**0.0856**	1.030	**0.0631**	NA	NA
C2	0.9885	0.2342	0.9869	0.2771	NA	NA	0.9706	**0.0631**
C3	1.0016	0.7303	NA	NA	1.013	0.2771	0.9834	**0.0856**
C4	NA	NA	0.9984	0.7303	1.012	0.2342	0.9819	**0.0652**
Age	1.0283	**7.37e-5**	1.0283	**7.37e-5**	1.028	**7.37e-5**	1.0283	**7.37e-5**

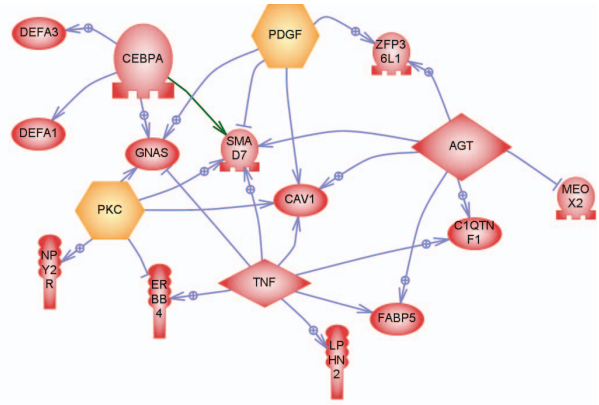


Fig. 6. Subnetwork enrichment analysis for Subtype 1 reveals AGT, PDGF, PKC, TNF, and CEBPA as dominant regulators with p-value of less than 0.05.

to repair the damage by expressing AGT. In GBM, AGT maintains normal function of vasculature [13] and cellular concentration of this enzyme is a primary determinant of the cytotoxicity of TMZ [14] in vitro. PKC (Protein Kinase C) is well established in cancer signaling and therapy as it is involved in proliferation, migration, and malignant transformation [15], and its isozyme has been suggested for chemotherapeutic targets in GBM [16]. TNF refers to a group of cytokines that induce proliferation, and inflammation and apoptosis depending upon the adaptor proteins. TNF is part of the anti-tumor strategy in which human glioma cell lines express its proteins. Manipulation of these proteins has shown to induce apoptosis in glioma cells [17]. Other hubs are highly ranked in the TCGA gene tracker.

4. CONCLUSION

In this paper, we proposed a multilevel framework for summarization of histological data and subsequent integrated analysis. Instead of directly summarizing cellular features at the patient level, we introduced an intermediate analysis level that summarizes cellular features within a local neighborhood and provide a compositional representation as patient level summarization. We then applied the proposed framework to TCGA Glioblastoma multiforme data for an integrated analysis. Our analysis indicates that one of the computed subtypes is prognostic and the molecular drivers, that correlate with compositional analysis of this subtype, are consistent with those in the GBM literature.

5. REFERENCES

[1] H. Chang, G.V. Fontenay, J. Han, G. Cong, F.L. Baehner, J.W. Gray, P.T. Spellman, and B. Parvin, "Morphometric analysis of tcga glioblastoma multiforme," *BMC Bioinformatics*, vol. 12, no. 484, 2011.

[2] H. Chang, L.A. Loss, P.T. Spellman, A. Borowsky, and B. Parvin, "Batch-invariant nuclear segmentation in whole mount tissue histology," in *Proc. ISBI*, 2012.

[3] L.A. Cooper et. al, "An integrative approach for in silico glioma research," *IEEE Trans Biomed Eng*, vol. 57, no. 10, pp. 2617–2621, 2010.

[4] L.A. Cooper et. al, "Integrated morphologic analysis for the identification and characterization of disease subtypes," *J Am Med Inform Assoc*, 2012, In Press.

[5] S. Monti, P. Tamayo, J. Mesirov, and T.R. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Mach Learn*, vol. 52, pp. 91–118, 2003.

[6] P.J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, pp. 53–65, 1987.

[7] R.G. Verhaak et. al, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[8] J. Fox, *Cox proportional-hazard regression for survival data, In R. Fox (Ed.), An R and S-PLUS companion to applied regression*, Sage, Thousand Oaks, CA, 2002.

[9] Y. Benjaminiand Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc Series B*, vol. 57, pp. 289–300, 1995.

[10] A. Pierallini, M. Bonamini, P. Pantano, F. Palmeggiani, M. Raguso, M.F. Osti, , G. Anaveri, and L. Bozzao, "Radiological assessment of necrosis in glioblastoma: variability and prognostic value," *Neuroradiology*, vol. 40, no. 3, pp. 150–153, 1998.

[11] Y. Liu, C. Li, and J. Lin, "Stat3 as a therapeutic target for glioblastoma," *Anticancer agents Med Chem*, vol. 10, no. 7, pp. 512–519, 2010.

[12] S.Rahman, P. Harbor, O. Chernova, G. Barnett, M. Vogelbaum, and S. Haque, "Inhinbition of constitutively active stat3 supresses proliferation and induces apoptosis in glioblastoma multiforme cells," *Oncogene*, vol. 21, no. 55, pp. 8404–8413, 2002.

[13] Y. Kakinuma, H. Hama, F. Syugiyama, K. Goto, K. Murakami, and A. Fukamizu, "Anti-apoptotic action of angiotensin fragments to neuronal cells from angiotensinogen knock-out mice," *Neuroscience Letters*, vol. 232, pp. 167–170, 1997.

[14] R. Stupp, M. Gander, S. Leyvraz, and E. Newland, "Current and future development in the use of temozolomide for the treatment of brain tumours," *The Lancent Oncology*, vol. 2, pp. 552–560, 2001.

[15] M. Kazanietz, *Protein Kinase C in cancer signaling and therapy*, Humana Press, 2010.

[16] P. Martin and I. JHussanini, "Pkc eta as a therapeutic target in glioblastoma multiforme," *Expert Opin Ther Targets*, vol. 9, no. 2, pp. 299–313, 2005.

[17] T. Chen, D. Hinton, B. Sippy, and F. Hoffman, "Soluable tnf-alpha receptors are constitutively shed and down regulate adhesion molecule expression in malignant gliomas," *Neuropathology*, vol. 56, pp. 541–550, 1997.