Morphometic Analysis of TCGA Glioblastoma Multiforme

Hang Chang[1], Gerald V. Fontenay[1], Ju Han[1], Ge Cong[1], Frederick L. Baehner[2], Joe W. Gray[1,3], Paul T. Spellman[1,3], and Bahram Parvin[1*]
[1] Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, [2] Department of Pathology, University of California, San Francisco, San Francisco, California, United States of America, [3] Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon, United States of America, *
Corresponding author: b_parvin@lbl.gov

Abstract

Background: Our goals are to develop a computational histopathology pipeline for characterizing tumor types that are being generated by The Cancer Genome Atlas (TCGA) for genomic association. TCGA is a national collaborative program where different tumor types are being collected, and each tumor is being characterized using a variety of genome-wide platforms. Here, we have developed a tumor-centric analytical pipeline to process tissue sections stained with hematoxylin and eosin (H&E) for visualization and cell-by-cell quantitative analysis. Thus far, analysis is limited to Glioblastoma Multiforme (GBM) and kidney renal clear cell carcinoma tissue sections. The final results are being distributed for subtyping and linking the histology sections to the genomic data.

Results: A computational pipeline has been designed to continuously update a local image database, with limited clinical information, from an NIH repository. Each image is partitioned into blocks, where each cell in the block is characterized through a multidimensional representation (e.g., nuclear size, cellularity). A subset of morphometric indices, representing potential underlying biological processes, can then be selected for subtyping and genomic association. Simultaneously, these subtypes can also be predictive of the outcome as a result of clinical treatments. Using the cellularity index and nuclear size, the computational pipeline has revealed five subtypes, and one subtype, corresponding to the extreme high cellularity, has shown to be a predictor of survival as a result of a more aggressive therapeutic regime. Further association of this subtype with the corresponding gene expression data has identified enrichment of (i) the immune response and AP-1 signaling pathways, and (ii) IFNG, TGFB1, PKC, Cytokine, and MAPK14 hubs.

Conclusion: While subtyping is often performed with genome-wide molecular data, we have shown that it can also be applied to categorizing histology sections. Accordingly, we have identified a subtype that is a predictor of the outcome as a result of a therapeutic regime. Computed representation has become publicly available through our Web site.

Background

While molecular characterization provides average genome-wide profiling for each biopsy, it fails to reveal inherent heterogeneity that is only visible through tissue histology. Molecular characterization has the advantage of a standardized array-based measurement compared to the genome and other well curated databases. On the other hand, histology sections do not provide standardized measurements, yet they are rich in content and continue to be the gold standard for the assessment of tissue neoplasm. Because of inter- and intra- observer variations [1] and the absence of quantitative representation, some studies have leveraged genome-wide analysis for improved markers for predicting biological behavior. If hematoxylin and eosin (H&E) stained tissue sections can be characterized in terms of cell type (e.g., epithelial, stromal), tumor type, and histopathological descriptors (e.g., tumor specific necrotic rate), then a richer description can be linked with genomic information for an improved basis for diagnostic and therapy. This is the main value of histological imaging since it captures detailed

morphometric features on a cell-by-cell basis and their organization. We have tested our system on Glioblastoma Multiforme (GBM), one of the most common and the least curable brain cancer, with glioma cells infiltrating the surrounding tissue with a median survival rate of 14.6 month [2]. Figure 1 shows that the tissue section has a rich spatial composition (e.g., lymphocytes in the lower right side, tumor cells), which is lost through bulk genome-wide array analysis (e.g., microarray, copy number). Our goal is to identify morphometric subtypes, based on nuclear structure and organization, from a very large sample size. First, we provide a brief review of the current state of art and then proceed with the details of our computational strategy. Present techniques for morphometric analysis have focused on several different aspects of tissue characterization, and they are summarized below along with a review of the nuclear segmentation from the H&E sections.

Brief review of analysis of H&E images: A comprehensive review of techniques for the analysis of the H&E sections is beyond the scope of this paper. However, a brief review can be found in [3]. From our perspective, three key concepts have been introduced to establish the trend and direction of the research community: **(I)** one group of researchers have focused on tumor grading through either accurate or rough nuclear segmentation [4] followed by computing cellular organization [5, 6] and classification. In some cases, tumor grading has been associated with recurrence, progression, and invasion carcinoma (e.g., breast DCIS) [7], but such an association is highly dependent on tumor heterogeneity and mixed grading (e.g., presence of more than one grade), which offers significant challenge to the pathologists as mixed grading appears to be present in 50% of patients [8]. A recent study indicates that detailed segmentation and multivariate representation of nuclear features from H&E stained sections can predict DCIS recurrence [9, 10] in patients with more than one nuclear grade. In this study, nuclei in the H&E stained samples were manually segmented and a multidimensional representation was computed for differential analysis between the cohorts. The significance of this particular study is that it has been repeated with the same quantitative outcome. In other related studies, image analysis of nuclear features has been found to provide quantitative information that can contribute to diagnosis and prognosis values for carcinoma of the breast [11, 12], prostate [13], and colorectal mucosa [14]. **(II)** The second group of researchers have focused on patch-based (e.g., region-based) analysis of tissue sections through means of supervised classification. These methods operate by representing each patch with color and texture features [15, 16] for training either a kernel or regression tree classifiers. A recent study evaluated and compared emerging techniques of sparse coding with kernel based methods (e.g., support vector machine, kernel discriminant analysis) on a GBM dataset to conclude that the kernel based method did equally as well, if not better, than sparse coding. Alternatively, some researchers have investigated how architectural features of tumor grades correlate with fractal dimensions [17]. Fractal dimensions differ from topological dimensions and has been shown to have the potential to elucidate irregularities by assigning a gross scalar value for discriminating benign and malignant breast cells from fine needle aspiration [18]. **(III)** A third group of researchers have suggested utilizing the detection of lymphocytes as a prognostic tool for breast cancer [19]. Lymphocytes are part of the adaptive immune response and their presence has been correlated with nodal metastasis and HER2-positive breast cancer, ovarian [20], and GBM. These cells often respond in larger quantity, and they can be easily detected because of their constant size (e.g., approximately 7 micron in diameter) and high chromatin content.

Brief review of methods for nuclear segmentation: Complexities in delineating nuclear regions originate from both technical (e.g., non-uniform fixation and staining protocol, artifacts in a tissue section, non-uniform thickness in tissue sections) and biological (e.g., different cell types, overlapping compartment) variations. Present techniques have focused on adaptive thresholding followed by morphological operators [21, 22], fuzzy clustering [4, 23], level set method using gradient information [24-26], color separation followed by optimum thresholding and learning [27, 28], hybrid color and texture analysis that are followed by learning and

unsupervised clustering [29], and representation of nuclei organization in tissue [30, 31] that is computed from either interactive segmentation or a combination of intensity, texture, and morphological operators. Some applications combine the above techniques. For example, in [32], iterative radial voting [33] was used to estimate seeds for the location of the nuclei and subsequently, the model interaction between neighboring nuclei with multiphase level set [34, 35]. It is also a common practice that through color decomposition, nuclear regions can be segmented using the same techniques that have been developed for fluorescence microscopy. In recent papers, we [36, 37] and others [38] have reviewed those techniques. However, none of these methods can effectively address analytical requirements of the tumor characterization. Thresholding and clustering assume constant chromatin content for the nuclei in the image. In practice, there is a wide variation in chromatin content. In addition, there is the issue with overlapping and clumping of the nuclei, and sometimes, due to the tissue thickness, they cannot be segmented. The method proposed in [32] aims to delineate overlapping nuclei through iterative radial voting [33], but seed detection can fail in the presence of wide variations in the nuclear size; thus, leading to fragmentation. We should also note that many of the techniques that have been developed for analysis of cell culture models, imaged through fluorescence microscopy, are applicable to the analysis of histology sections. Accordingly, methods have been developed to quantify a variety of endpoints using iterative voting [33, 39], geometric reasoning [40, 41], evolving fronts [35, 37, 42], and Gabor filter banks [43].

Having summarized the current state of computational histopathology, our objective is to use a large growing dataset of tumor sections and to identify intrinsic subtypes within this dataset. These subtypes can then be used for genomic association. In other words, we don't seek to build a system to mimic histological grading. To meet this objective, it is essential to develop a pipeline for processing a large scale dataset, to overcome technical variations, and to incorporate methods that are extensible to other tumor types. Our testbed consists of 344 sections of GBM, scanned with a 20X objective in a bright field, which are typically 40,000-by-40,000 pixels.

Method

Morphometric analysis and multidimensional profiling: We evaluated a number of nuclear segmentation methods that included level sets [44] or their variants using graph cut implementation, and integration of these methods with seed selection using geometric methods [42]. But these techniques proved to be compute-intensive as a typical tissue section (of size 40k-by-40k pixels) would take roughly a week of processing time on a high end desktop computer. Our experience led to a design of a pipeline that will delineate nuclei and compute morphometric features with a superior computational throughput. The computational model was first validated against synthetic data, then tested on annotated tissue sections, and finally evaluated by a pathologist. Below, we summarize three major components of our methodology.

Analytical steps: Figures 2 and 3 show steps in converting an image into a multidimensional representation. (I) The first step removes heterogeneity associated with staining by normalizing against one gold standard of H&E stain. (II) The second step performs color decomposition for further reduction of the computational load. The standard approach is a non-negative matrix factorization (NMF) [45], but it is iterative and a previous analysis has indicated NMF did not show superior performance [28]. Here, we used a linear transformation for separating stains [46] based on the orthonormal transformation of the RGB space. (III) The third step computes a threshold from the image corresponding to the nuclei signature. The threshold selection is based on the analysis of the histogram for the value that minimizes intra-class variance. Other techniques, such as modeling foreground and background as two Poisson distributions, yielded similar results. The important issue is fast histogram-based thresholding for subsequent refinement and validation. Refinement consists of enforcement of intensity and geometric

constraints. Often, when nuclei are close to each other, either their intercellular contents can leak, the boundaries between the two adjacent nuclei can become perceptual, or the two neighboring nuclei, with completely different chromaticity strength, can merge. The refinement step performs two tasks: (i) it models the intensity distribution of each thresholded blob as a mixture of up to three Gaussians to examine if there is a variation in the background model and whether two adjacent nuclei, with a significantly different amount of DNA content, are merged together, and (ii) it uses the convexity constrain to partition blobs based on perceptual boundaries, as outlined in our earlier paper [40]. Once an image is segmented in terms of nuclear morphology, a multidimensional representation is generated for each nucleus that defines its signature and organization, as we defined in a previous publication [37] and summarized in Additional file 1.

Computational pipeline: The significance of the pipeline, shown in Figure 4, is that it can process a large amount of data; thus, meeting TCGA data processing requirements. The pipeline has four components: (I) maintaining consistency between the remote and local registries, (II) visualization of tissue sections, (III) data processing and importing computed representation, and (IV) data summarization through normalization.

(I) The pipeline maintains a local registry where consistency between images at TCGA (at the National Cancer Institute) and a local repository is constantly maintained, and new images are downloaded for processing. At present, NCI provides both frozen sections and those from paraffin embedded blocks. Although both types of images are registered and displayed through our system, only paraffin embedded blocks are processed. Each image is partitioned into strips of 1k-by-number of columns, then the strips are stored on the OME image server [47, 48].

(II) Visualization of each large scale tissue section is realized through tiling and the utilization of Flash technology that enables a client to pan and zoom, similar to *GoogleMaps™*. Each image (of the order of 40k-by-40k pixels or higher) is partitioned into tiles of 256-by-256 pixels at different resolutions, and the tiles are then stored on a server. As the user drags and zooms on the image in the browser, the tiles are downloaded from the server and inserted into the browser page. Data and images are available through http://tcga.lbl.gov.

(III) Each strip is subsequently partitioned into 1k-by-1k blocks, and blocks are submitted to a computer cluster for processing. The block size has been optimized for processing time and wait time in the queue. At the moment, the entire GBM data set of 344 images takes 4 days of processing. In addition to cluster-based computing, the computational methods of the previous section have a multithread implementation for a more efficient utilization of each computing node. Once each block is processed, computed features are imported into an imaging bioinformatics system, named BioSig [37, 49], for further analysis. Several java modules have been developed that run concurrently to access and update the database. The “Jobsubmitter” uses *JSch* (java version of ssh), and *ExpectJ* (java version of Expect) to drive shell scripts on the computing cluster. Computed representation (e.g., nuclear segmentation) can then be overlaid on the original image for quality control.

(IV) The backend of BioSig uses PostgreSQL (PG) and summarization of feature-based representation is performed through procedural programming. For high performance applications, PG server programming interface (SPI) enables the transparent transformation of SQL queries. This is a critical component since it adds flexibility for computing morphometric and organization features, normalizing them, and analyzing underlying representation in a new way that was not anticipated. This capability has proven to increase productivity by testing alternative representations without reprocessing the original images. Given the entire GBM (or other tumors) datasets, we have designed a four-step process to normalize each computed

feature (e.g., nuclear size, texture, cellularity) for subtyping and genomic association, which is implemented through SPI: (i) each feature is represented as a density distribution per tissue; (ii) feature-based distribution for all tissues, within a tumor type, are combined to construct a global distribution; (iii) the global distribution is then re-binned so that each bin has a similar population of cells of a given feature-value, and (iv) local density distributions are then remapped to computed global bins of equal weight. The net result is that the morphological indices can then be compared, in context, by reporting a distribution function for each feature. These data are downloadable and can be visualized for each tissue section. The rationale for this simplified analysis is that given a large number of cells in a tissue section, classical clustering analysis (for quantization) can be computationally intractable (e.g., computing similarity matrices). In cases where multiple tissue sections exist for a single patient, an average distribution is computed and archived.

Subtyping and genomic association: Normalized representation of morphometric data are used for subtyping. Subtyping is based on consensus voting [50] by varying the number of subtypes and examining the similarity matrix. It has also been used in earlier papers for subtyping 2D and 3D cell culture morphologies [43, 51]. Two gene ranking algorithms of moderated F-statistic and random forests are used for genomic association. (i) Moderated F-statistic [52] utilizes the empirical Bayes method for assessing differential gene expression. In this method, the denominator mean squares (e.g., variance) are moderated across genes through the empirical Bayes approach. The net result is an improved statistical stability given the limited number of samples. The p-value is computed for each gene based on the moderated F-statistic, and then adjusted for multiple hypothesis testing. The adjustment is based on Benjamini and Hochberg's method to estimate the false discovery rate (FDR) [53]. FDR controls the expected proportion of falsely rejected null hypotheses in multiple hypotheses testing to correct for multiple comparisons. The method is implemented through the R Limma package. The top genes that are differentially expressed between subtype 5 and others, with FDR adjusted p-value less than 0.06, are included in Additional file 1 as a heatmap. (ii) Random forest is an ensemble classifier that consists of many decision trees [54]. In random forest, there are several policies for characterizing significance of each gene. One policy evaluates the decrease in classification accuracy by permutation values of a single gene between multiple samples [55]. We used the R implementation of a random forest package [56], where the number of trees (ntree) is increased to 2000 to accommodate the original subset of genes (1740) that were used in an earlier TCGA publication [57]. To insure the robustness and stability of gene selection, the process is repeated by averaging over 100 randomly generated forests.

Results

The critical factors in our computational pipeline are the throughput, quality of segmentation and morphometric representation for subtyping, and genomic association. The throughput is significant since images need to be continuously processed with a newer version of the software with increased robustness. Presently, the total computational time for 344 large scale tissue sections (from 133 patients) is less than a week on a shared cluster. Because segmentation results are also important for quality control, a number of intermediate data are also released.

Data, intermediaries, and limitations: Since nuclear segmentation provides the basis for morphometric analysis, subtyping, and survival analysis, it is being released for visualization through our web site at http://tcga.lbl.gov, where users can pan and zoom through the images and overlay segmentation results on original images. The web site also enables exclusion of specific tissue sections for subtyping and genomic association. Computed representations and subtyping is also released through our web site to the community.

Present limitations are absence of (i) an improved nuclear segmentation method, (ii) patch-based tissue-based labeling, (iii) a systematic evaluation of the multidimensional representation,

as it relates to the underlying biology, and (iv) abstraction and complete automation in the computational pipeline. (i) Like others, our approach to nuclear segmentation is not perfect and introduces morphometric errors. The major limitation for introducing more powerful algorithms has been limited computational time for processing very large sets of data. However, given a very large number of cells in a tissue section, subsequent consensus-based clustering tends to treat segmentation errors as outliers. Figure 5 shows nuclear segmentation and region-based tessellation overlaid on images with diverse morphometric signatures, where the cellularity index (e.g., density of cells in a region) is computed as the inverse of each tessellated region area and its density distribution. (ii) In certain tumor types, nuclear segmentation is insufficient for characterizing tissue histopathology. For example, in GBM, apoptotic and necrotic rates are also important. There are also patches where the state of the tissue is transitory, i.e., both apoptotic and necrotic states coexist in a population of cells. This is a higher level of analysis that is difficult to deduce from a simple nuclear segmentation and additional prior knowledge is needed. (iii) Over 50 features are computed per cell, and we have only begun to evaluate some of those that correlate with the known pathology (e.g., nuclear size, cellularity). It is desirable to have an informatics layer for formulating a query and get a different view of the data. Possible use-cases are dimensionality reduction (e.g., PCA, MDS), or feature selection based on outcome or known pathology that is followed by subtyping. Each of these queries provide a unique insight and into the underlying biology for hypothesis generation. (iv) Ideally, all processes should be launched, monitored, and validated through the database. Although, images and computed features are registered with the database, additional queries and notification services are required to construct a more flexible system as required in items (ii) and (iii).

Quality control: Three modules are tested in the computational pipeline: (i) segmentation, (ii) feature extraction, and (iii) subtyping. (i) We have created a subset of hand segmented images, which originate from a diverse set of tissue sections from TCGA GBM dataset. Even though most images are stained properly, the emphasis on this subset has been placed on blocks where the nuclear dye is heterogeneous. The recall and precision is at 78% and 65%, respectively; (ii) feature extraction and representation were tested against synthetic data with known ground truth; (iii) subtyping is evaluated qualitatively by displaying group similarity matrix.

Subtyping based tissue histology and survival analysis: Our system represents each nucleus as a multidimensional vector in the tissue section. We have opted the policy to allow the pathologist to explore clinical questions in terms of selected morphometric indices. This is based on the fact that each morphometric feature can represent underlying biological processes. For example, when the cells are stressed, macromolecules are excreted into cytoplasm (or ECM) to create a textured topography as opposed to a smooth one in normal cells. In the following experiment, it was decided to investigate nuclear size and cellularity for subtyping, survival analysis, and genomic association. The rationale is clear given larger nuclear size and higher proliferation rate in tumor regions. In this experimental configuration, consensus voting revealed five subtypes through qualitative analysis and ordering of the computed similarity matrix, as shown in Figure 5. With respect to correlation with the outcome as a result of therapy, we analyzed patients that received more (e.g., concurrent radiation and chemotherapy or greater than 4 cycles of chemotherapy) or less (e.g., non-concurrent radiation and chemotherapy or less than 4 cycles of chemotherapy) [57]. Following the Kaplan Meier estimator, our analysis indicates that only one subtype with extreme high cellularity, shown in Figure 5E, has a significant p-value through pair-wise comparison of the survival curves using a log-rank test [58]. Figure 6F indicates that with a more intensive therapy (the red curve) life span is increased as compared to a less intensive therapy (blue line). The p-values of other subtypes were not favorable for survival analysis. A possible interpretation is that extreme high cellularity is more homogeneous and highly proliferative; thus, responding better to a more aggressive therapy.

Genomic association: Given that the therapeutic regime has increased life span for the subtype with extreme high cellularity, as shown in Figure 5E, we queried for its molecular marker through differential gene expression analysis as well as random forest. Both gene lists are provided in Additional file 1, and a more detailed discussion of the gene lists through random forest follows. We have analyzed the top 100 genes for pathway and subnetwork enrichment analysis through *Pathway Studio*. Pathway analysis reveals enrichment of immune response, such as NK-cell (Natural killer cell) and T-cell activation, and AP-1 signaling with p-value of less than 0.05. In support of these findings, the literature suggests that GBM expresses antigens that is recognized by the immune system to eliminate virus infected cells and GBMs [59, 60]. Tumor associated antigen (TAA) indicates that glioma cells can be recognized by the immune response, but this process is hindered by the tumor location and evasion strategies developed by GBM. AP-1 (JUN oncogene) is a transcription factor is responsible for high level regulation of IL-13Ra2 that is expressed in GBM cells [61], and is also a highly ranked gene in the TCGA gene tracker.

Subnetwork enrichment analysis has revealed six hubs, with p-values of less than 0.05 that regulate eight or more other components. These are *IFNG, TGFB1, MAPK14*, Cytokines, *PKC,* and *IL1B*. The union of these subnetworks is shown in Figure 7. *IFNG* and *MAPK14* are shown to be highly scored by TCGA gene tracker; *TGFB1* is known to be upregulated in GBM [62]; *PKC* (Protein Kinase C) is well established in cancer signaling and therapy as it is involved in proliferation, migration, and malignant transformation [63], and its isozyme has been suggested for chemotherapeutic targets in GBM [64]; and *ILB1* is down stream of *NF-kB* and is known to play an important role in cellular response to stress [65] and is constitutively activated in most tumor types. In summary, bioinformatics analysis has provided hypotheses for new modes of potential therapy based on morphometric subtyping.

Comparison with prior art: It is important to note that another laboratory [66] has analyzed the same dataset. There are difference in the outcome and methodologies. For example, they have reported four subtypes in the GBM dataset. We suggest the (i) addition of the cellularity index, (ii) utility of feature distributions as opposed to the feature means, (iii) selection of specific combination of features as opposed to all computed features, and (iv) absence of curation have been the deciding factors. Besides cell-based multidimensional representation, there are also differences in nuclear segmentation. It is difficult to assess the differences in segmentation in the absence of source code and computed results on a large dataset; however, color normalization (with respect to the gold standard) and separation of the touching nuclei has not been addressed in [66]. These differences, especially curation, can have a significant impact on morphometric analysis. Finally, we have designed and built an open system, where algorithms and software are going through constant improvement, and computed representation and intermediaries are being made available for each version of the software.

Conclusions

We have developed an integrated pipeline to process large scale tissue sections for morphometric analysis. The data are downloaded from the NIH web site, partitioned into blocks, and then processed on a cluster. Computed representation is then transferred to a database where (i) data can be downloaded for molecular association, and (ii) computed information is overlaid on the original image and that through panning and zooming, quality control can be performed. Thus far, GBM and kidney data have become publicly available.

We have shown that through morphometric analysis and cellular organization of tissue histology of a large dataset, subtypes can be identified that are predictive of outcome as a result of therapeutic protocol. The main theme is that histological subtyping reveals intrinsic categories that are independent of supervised histological grades. In other words, TCGA's large curated dataset offers potential for revealing subtypes based on *intrinsic* properties of tissue signatures

as opposed to the classical tumor grading (e.g., Gleason ranking in prostate cancer), practiced by pathologists. In this context, TCGA's histology database can provide a complementary repository for diagnostic and molecular underpinning for histological subtypes. Subsequently, molecular signature of a subtype can hypothesize a more effective targeted therapy. Our continued research focuses on addressing limitations that has been addressed in the Result section. Ultimately, we plan to develop a system that will process all tumor types.

Acknowledgements

This research was supported by the National Cancer Institute U24 CA1437991 and RO1CA140663 and the U.S. Department of Energy under contract number DE-AC02-05CH11231. The research utilized the Lawrencium computational cluster resource, provided by the IT Division at the Lawrence Berkeley National Laboratory, and the Center for Information Technology Research in the Interest of Society (CITRIS) at the University of California-Berkeley.

Authors' contribution

HC developed high performance software for segmentation and feature-based representation. GVF developed integrated computational pipeline for routing data between NIH repository, local databases, and cluster computing. JH performed subtyping and bioinformatics analysis of multidimensional representation. GC developed the web pages for data distribution and image zooming. FB is the lead pathologist who contributed to the interpretation of subtyping results. JG and PS conceived the study and requirements. BP contributed to the study and experimental plan, led the research, and wrote the paper. All authors read and approved the final manuscript.

References

1. Dalton L, Pinder S, Elston C, Ellis I, Page D, Dupont W, Blamey R: **Histolgical gradings of breast cancer: linkage of patient outcome with level of pathologist agreements** *Modern Pathology* 2000, **13**:730-735.
2. Stupp R, Mason W, vanen Bent M, Weller M, Fisher B, Taphoorn M, Belanger K, Brandes A, Marosi C, Bogdahn U *et al*: **Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma**. *New England Journal of Medicine* 2005, **352**(10):987-996.
3. Demir C, Yener B: **Automated cancer diagnosis based on histopathological images: a systematic survey**. In.: Rensselaer Polytechnic Institute; 2009.
4. Latson L, Sebek N, Powell K: **Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy**. *Analytical and Quantitative Cytology and Histology* 2003, **26**(6):321-331.
5. Doyle S, Feldman M, Tomaszewski J, Shih N, Madabhushu A: **Cascade multi-class pairwise classifier (CASCAMPA) for normal, cancerous, and cancer confunder classes in prostate histology**. In: *International Synposium on Biomedical Imaging: from nano to macro* IEEE; 2011: 715-718.
6. Basavanhally A, Xu J, Madabhushu A, Ganesan S: **Computer-aided prognosis of ER+ breast cancer histopathology and correlating survival outcome with oncotype DX assay**. In: *International Synposium on Biomedical Imaging: from nano to macro* IEEE; 2009: 851-854.
7. Kerlikowske K, Molinaro A, Cha I, al e: **Characteristics associated with recurrence among women with ductal carcinoma in situ treated by lmpectomy**. *Journal of the National Cancer Institute* 2003, **95**:1692-1702.
8. Miller N, Chapman J, Fish E: **In situ duct carcinoma of the breast: clinical and histopathologic factors and association with recurrent carcinoma**. *Breast Journal* 2001, **7**:292-302.

9. Axelrod D, Miller N, Lickley H, Qian J, Christens-Barry W, Yuan Y, Fu Y, Chapman J: **Effect of quantitative nuclear features on recurrence of ductal carcinoma *in situ* (DCIS) of breast**. *Cancer Informatics* 2008, **4**:99-109.
10. Chapman J, Miller N, Lickley H, Qian J, Christens-Barry W, Fu Y, Yuan Y, Axelrod D: **Ductal carcinoma in situ of the breast (DCIS) with heterogeneity of nuclear grade: prognostic effects of quantitative nuclear assessment**. *BMC Cancer* 2007, **7**(174).
11. Peinta K, Coffey D: **Correlation of nuclear morphometry with progression of breast cancer**. *Cancer* 1991, **68**:2012-2016.
12. Mommers E, Poulin N, Sangulin J, Meiher C, Baak J, van Diest P: **Nuclear cytometric changes in breast carcinogenesis**. *Journal of Pathology* 2001, **193**(1):33-39.
13. Veltri R, Khan M, Miller M, Epstein J, Mangold L, Walsh P, Partin A: **Ability to predict metastasis based on pathology findings and alterations in nuclear structure of normal appearing and cancer peripheral zone epithelium in the prostate**. *Clinical Cancer Research* 2004, **10**:3465-3473.
14. Verhest A, Kiss R, d'Olne D, Larsimont D, Salman I, de Launoit Y, Fourneau C, Pastells J, Pector J: **Characterization of human colorectal mucosa, polyps, and cancers by means of computerized mophonuclear image analysis**. *Cancer* 1990, **65**:2047-2054.
15. Bhagavatula R, Fickus M, Kelly W, Guo C, Ozolek J, Castro C, Kovacevic J: **Automatic identification and delineation of germ layer components in H&E stained images of teratomas derived from human and nonhuman primate embryonic stem cells**. In: *International Synposium on Biomedical Imaging: from nano to macro* IEEE; 2010: 1041-1044.
16. Kong J, Cooper L, Sharma A, Kurk T, Brat D, Saltz J: **Texture based image recognition in microscopy images of diffuse gliomas with multi-class gentle boosting mechanism**. In: *ICASSAP: 2010*; 2010: 457-460.
17. Tambasco M, Magliocco A: **Relationship between tumor grade and computed architectural complexity in breast cancer specimens**. *Human Pathology* 2008, **39**(5):740-746.
18. Dey P, Mohanty S: **Fractal dimensions of breast lesions on cytology smears**. *Diagn Cytopathol* 2003, **29**:85-87.
19. Fatakdawala H, Xu J, Basavanhally A, Bhanot G, Ganesan S, Feldman F, Tomaszewski J, Madabhushi A: **Expectation-maximization-driven geodesic active contours with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology**. *IEEE Transactions on Biomedical Engineering* 2010, **57**(7):1676-1690.
20. Zhang L, Conejo-Garcia J, Katsaros P, Gimotty P, Massobrio M, Regnani G, Makrigiannakis A, Gray H, Schlienger K, Liebman M *et al*: **Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer**. *New England Journal of Medicine* 2003, **348**(3):203-213.
21. Phukpattaranont P, Boonyaphiphat P: **Color based segmentation of nuclear stained breast cancer cell images**. *ECTI Transactions on Electrical Engineering, and Communication* 2007, **5** (2):158-164.
22. Ballaro B, Florena A, Franco V, Tegolo D, Tripodo C, Valenti C: **An automated image analysis methodology for classifying megakaryocytes in chronic myelprliferative disorders**. *Medical Image Analysis* 2008, **12**:703-712.
23. Land W, McKee D, Zhukov T, Song D, Qian W: **A kernelised fuzzy-Support Vector Machine CAD system for the diagnosis of lung cancer from tissue images**.

International Journal of Functional Informatics and Personalised Medicine 2008, **1**(1):26-52.

24. Bamford P, Lovell B: **Unsupervise cell segmentation with active contours**. *Signal Process* 1998, **71**(2):203-213.
25. Glotsos D, Spyridonos P, Cavouras D, Ravazoula P, Dadioti P, Nikiforidis G: **Automated segmentation of routinely hematoxyli-eosin stained microscopic images by combining support vector machine, clustering, and active contour models**. *Anal Quant Cytol Histol* 2004, **26**(6):331-340.
26. Fatakdawala H, Basavanhally A, Xu J, Bhanot G, Ganesan S, Feldman M, Tomaszewski J, Madabhushi A: **Expectation maximization driven geodesic active contour:application to lymphocyte segmentation on digitized breast cancer histhopatholgy**. In: *International conference on bioinformatics and bioengineering* 2009: 69-76.
27. Cosatto E, Miller M, Graf H, Meyer J: **Grading nuclear plemorphism on histological micrographs**. In: *International Conference on Pattern Recognition* 2008: 1-4.
28. Chang H, Defilippis RA, Tlsty TD, Parvin B: **Graphical methods for quantifying macromolecules through bright field imaging**. *Bioinformatics* 2009, **25**(8):1070-1075.
29. Datar M, Padfield D, Cline H: **Color and texture based segmentation of molecular pathology images usING HSOMS**. In: *International Symposium for Biomedical Imaging: from nano to maco* IEEE; 2008: 292-295.
30. Petushi S, Garcia F, Haber M, Katsinis C, Tozeren A: **Large-scale computations on histology images reveal grade-differentiation parameters for breast cancer**. *BMC Medical Imaging* 2006, **6**(14).
31. Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski: **Automated grading of breast cancer histipathology using spectral clustering with textural and architectural image features**. In: *International Symposium on Biomedical Imaging: from nano to macro* 2008: 496-499.
32. Bunyak F, Hafiane A, Palanippan K: **Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level set**. *Advances in Experimental Medicine and Biology* 2011, **696**:413-242.
33. Parvin B, Yang Q, Han J, Change H, Rydberg B, Barcellos-Hoff MH: **Iterative voting for inference of structural saliency and localization of subcellular structures**. *IEEE Transactions on Image Processing* 2007, **16**(3).
34. Nath S, Palaniappan K, Bunyak F: **Cell segmentation using coupled level sets and graph-vertex coloring**. In: *Medical Image Computing and Computed-assisted Intervention-Miccai: 2006*; 2006: 101-108.
35. Chang H, Parvin B: **Multiphase level set for automated delineation of membrane-bound macromolecules**. In: *International Symposium for Biomedical Imaging: from nano to macro* IEEE; 2010: 165-168.
36. Han J, Chang H, Yang Q, Groesser T, Barcellos-Hoff M, Parvin B: **Multiscale iterative voting for differential analysis of stress response for 2D and 3D cell culture models**. *Journal of Microscopy* 2010, **241**(3):315-326.
37. Han J, Chang H, Andrarwewa K, Yaswen P, Barcellos-Hoff M, Parvin B: **Multidimensional profiling of cell surface proteins and nuclear markers**. *IEEE Transactions on Computational Biology and Bioinformatics* 2010, **7**(1):80-90.
38. Coelho L, Shariff A, Murphy R: **Nuclear Segmentation in Microscope Cell Images: A Hand-Segmented Dataset and Comparison of Algorithms**. In: *International Symposium on Biomedical Imaging: from nano to macro* IEEE; 2009: 690-693.

39. Loss L, Bebis G, Parvin B: **Iterative tensor voting for perceptual grouping if ill-defined curvilinear structures**. *IEEE Transactions on Medical Imaging* 2011, **30**(8):1503-1513.
40. Wen Q, Chang H, Parvin B: **A Delaunay triangulation approach for segmenting a clump of nuclei**. In: *International Synposium on Biomedical Imaging: from nano to macro* IEEE; 2009: 9-12.
41. Raman S, Maxwell C, Barcellos-Hoff MH, Parvin B: **Geometric approach to segmentation and protein localization in cell culture assays**. *Journal of Microscopy* 2007, **225**(Part 1):22-30.
42. Chang H, Yang Q, Parvin B: **Segmentation of heterogeneous blob objects through voting and level set formulation**. *Pattern Recognition Letters* 2007, **28**(13):1781-1787.
43. Han J, Chang H, Giricz O, Lee G, Baehner F, Gray J, Bissell M, Kenny P, Parvin B: **Molecular Predictors of 3D Morphogenesis by Breast Cancer Cell Lines in 3D Culture**. *PLoS Computational Biology* 2010, **6**(2):e1000684.
44. Chan T, Vese L: **Active contours without edges**. *IEEE Transactions on Image Processing* 2001, **10**(2):266-277.
45. Rabinovich A, Agarwal S, Larris C, Price J, Belongie S: **Unsupervised color decomposition of histologically stained tissue samples**. In: *Advances in Neural Information Processing Systems* MIT Press; 2003: 667-674.
46. Ruifork A, Johnston D: **Quantification of histochemical staining by color decomposition**. *Anal Quant Cytol Histology* 2001, **23**(4):291-299.
47. Swedlow J, Goldberg I, Brauner E, Sorger P: **Informatics and quantitative analysis in biological imaging**. *Science* 2003, **300**:100-102.
48. Goldberg I, Allan C, Burel JM, Creager A, Falconi H, Hochheiser H, Johnston J, Mellen J, Sorger P, Swedlow J: **The open microscopy environment (OME) data model and xml files: open tools for informatics and quantitative analysis in biological images**. *Genome and Biology* 2005, **6**(5):R47.
49. Parvin B, Fontenay G, Yang Q, Barcellos-Hoff MH: **BioSig: an imaging bioinformatics system for phenotypic analysis**. *IEEE Transactions on System, Man, and Cybernetics-Part B* 2003, **33**(5):814-824.
50. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data**. *Machine Learning* 2003, **52**(1-2):91-118.
51. Han J, Chang H, Fontenay G, Wang N, Gray J, Parvin B: **Morphometric subtyping for a panel of breast cancer cell lines**. In: *International Symposium on Biomedical Imaging: from Nano to Macro* IEEE; 2009: 791-794.
52. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl GenetT Mo B* 2004, **3**:A3.
53. Benjamini Y, Hochberg, Y.: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J Roy Stat Soc B* 1995, **57**:397-420.
54. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
55. Diaz-Uriarte R, Alvarez de Andres, S.: **Gene selection and classification of microarray data using random forest**. *BMC Bioinformatics* 2006, **7**:3.
56. Liaw A, Wiener, M.: **Classification and Regression by randomForest**. *R News* 2002, **2**(3):18-22.
57. Verhaak R, Hoadley K, Purdom E, Wang V, Qi Y, Wilkerson M, Miller C, Ding L, Golub T, Mesirov J *et al*: **Integrated genomic analysis identifies clinically relevant**

subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 2010, **17**:98-110.
58. Mantel N: **Evaluation of survival data and two new rank order statistics arising in its consideration**. *Cancer Chemotherapy Reports* 1966, **50**(3):163-170.
59. Johnson L, Sampson J: **immunotherapy approaches for malignant glioma from 2007 to 2009**. *Curr Neurol Neurosi Rep* 2010, **10**(4):259-266.
60. Mitchell D, Sampson J: **Toward effective immunotherapy for the treatment of malignant brain tumors**. *Neurotherapeutics* 2009, **6**(3):527-538.
61. Wu A, Ericson K, Chao W, Low W: **NFAT and AP1 are essential for the expression of a glioblastoma multiforme related IL-13Ra2 transcript**. *Cell Oncology* 2010, **32**(5-6):313-329.
62. Lin B, Madan A, Yoon J, Fang X, Yan X, Kim T, Hwang D, Hood L, Foltz G: **Massively parallel signature sequencing and bioinformatics analysis identifies up-regulation of TGFBI and SOX4 in human glioblastoma**. *PLoS One* 2010, **5**(4).
63. Kazanietz M: **Protein Kinase C in cancer signaling and therapy**: Humana Press; 2010.
64. Martin P, JHussanini I: **PKC eta as a therapeutic target in glioblastoma multiforme**. *Expert Opin Ther Targets* 2005, **9**(2):299-313.
65. Keyse S: **Stress response: methods and protocols**. Totowa, New Jersey: Humana press; 2000.
66. Cooper L, Kong J, Wang F, Kurk T, Moreno C, Brat D, Saltz J: **Morphological Signatures and Genomic Correlates in Glioblastoma**. In: *International Symposium on Biomedical Imaging: from nano to macro* IEEE; 2011: 791-794.

Figure Legends

Figure 1: A pinhole view of GBM tumor section indicates a rich spatial composition in terms of nuclear size, cellularity, and presence of lymphocytes.

Figure 2: Steps in delineating each nucleus from an H&E stained tissue sections.

Figure 3: Steps in delineation of nuclei. (A) Reference image for color normalization, (B) Original H&E image, (C) normalized image, (D-E) color decomposition for each stain, (F) thresholding, and (G) refinement and validation.

Figure 4: Computational pipeline consists of four modules: downloads images from the NIH repository. Each image is partitioned into strips of (1k-by-number of columns), stored in the OMEIS image server. Each strip is partitioned into blocks of 1k-by-1k pixels, where each block is submitted to one of the two clusters at Berkeley Lab. Computed representations are then imported into a PostgreSQL database.

Figure 5: Nuclear segmentation and region-based tessellation for preferred subtypes of Figure 6E: (A) high cellularity, (B) low cellularity, (C) medium cellularity, (D) high cellularity with pleomorphism, and (E) extreme high cellularity.

Figure 6: Steps in identifying subtypes from morphological descriptors of a tissue section. (A) Each patient may have multiple tissue sections, which are accessible along with the computed features and coded clinical information through BioSig in (B). (C) Each feature, from each tissue, is represented as a density distribution that is normalized in (D). (E) Subtyping identifies 5 classes through consensus voting. (F) Following the Kaplan Meier test, only one subtype proved to have a significant p-value between pair-wise survival curves.

Figure 7: Subnetwork enrichment analysis has revealed 6 hubs with p-value<0.05: IFNG, TGFB1, MAPK14, Cytokine, PKC, and ILB1. Union of these subnetworks and interactions indicates interactions between of these hubs.

Additional files

Additional file 1
Title: Supplementary Material for Morphometic Analysis of TCGA Glioblastoma Multiforme.
Description: Supplementary Material for Morphometic Analysis of TCGA Glioblastoma Multiforme.

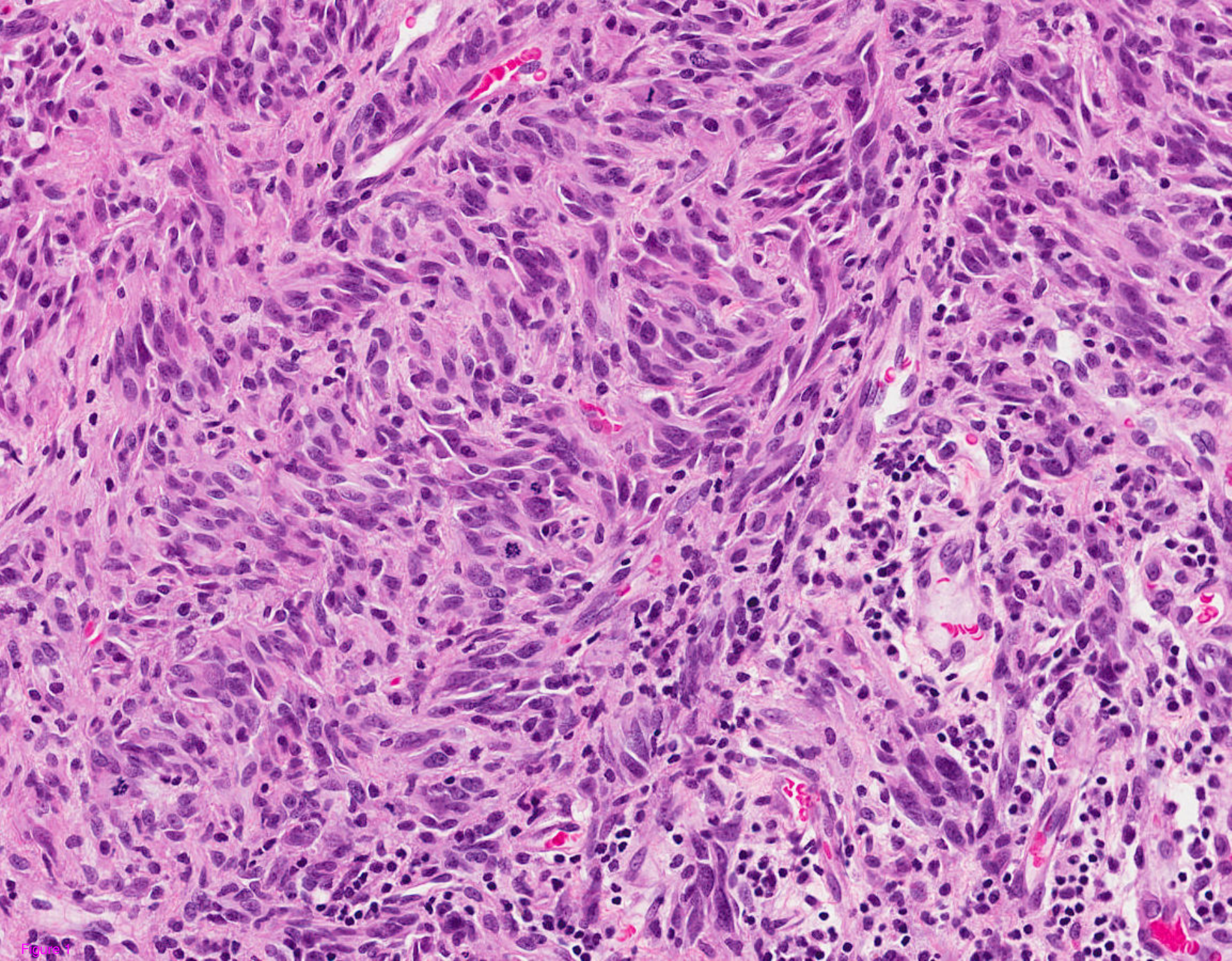


Figure 1

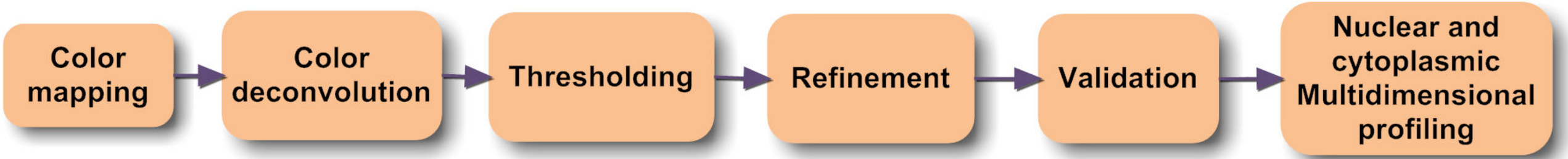

Color mapping
Color deconvolution
Thresholding
Refinement
Validation
Nuclear and cytoplasmic Multidimensional profiling

Figure 2

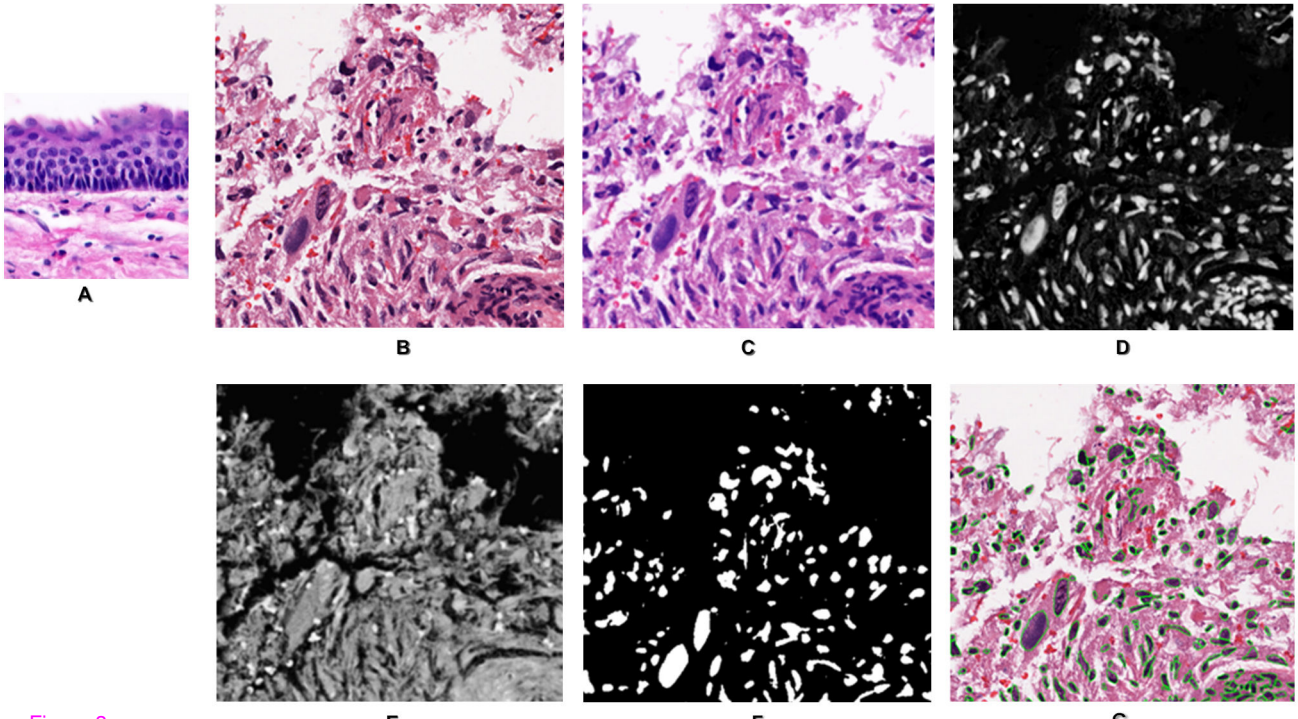


Figure 3

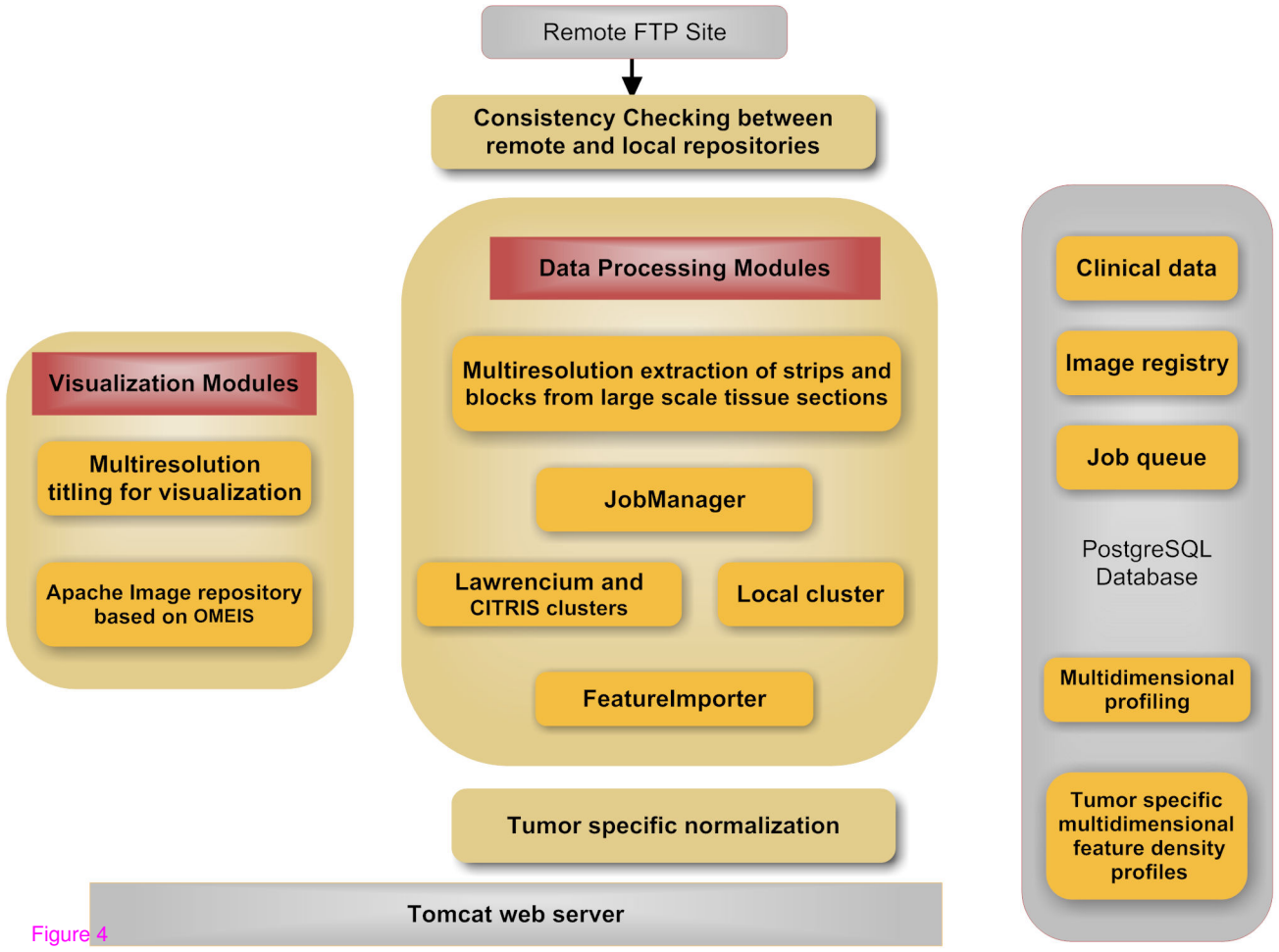
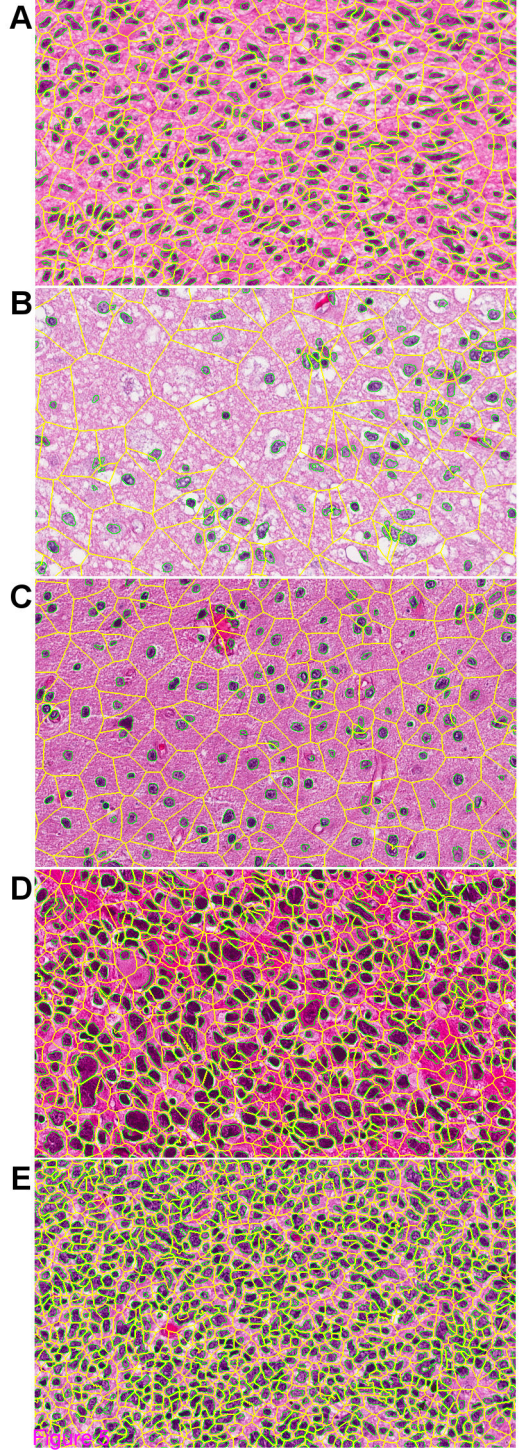

Remote FTP Site
Consistency Checking between remote and local repositories
Data Processing Modules
Multiresolution extraction of strips and blocks from large scale tissue sections
JobManager
Lawrencium and CITRIS clusters
Local cluster
FeatureImporter
Visualization Modules
Multiresolution titling for visualization
Apache Image repository based on OMEIS
Clinical data
Image registry
Job queue
PostgreSQL Database
Multidimensional profiling
Tumor specific multidimensional feature density profiles
Tumor specific normalization
Tomcat web server

Figure 4


A
B
C
D
E
Figure 2

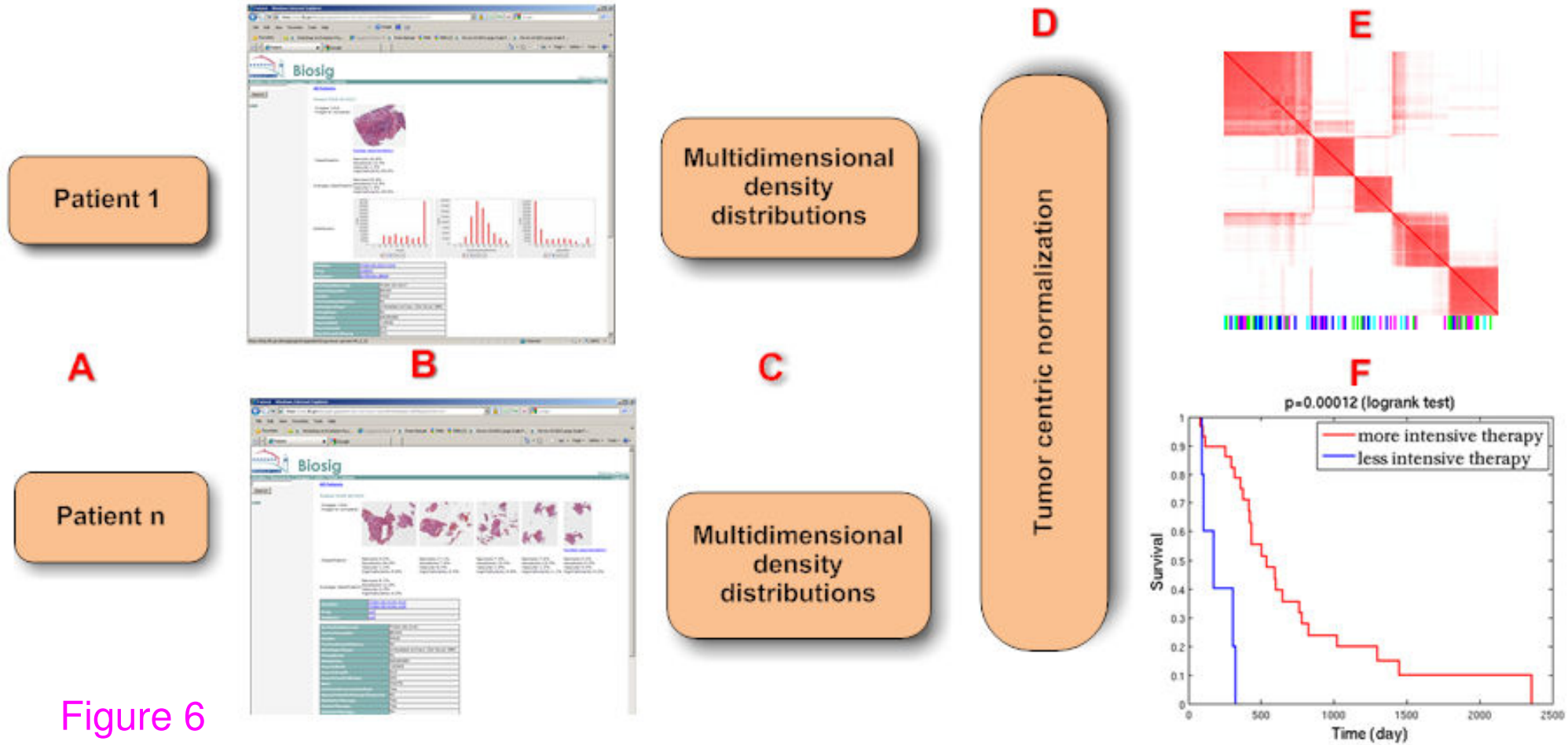

Patient 1
Patient n
A
B
C
D
E
F
Biosig
Biosig
Multidimensional density distributions
Multidimensional density distributions
Tumor centric normalization
p=0.00012 (logrank test)
more intensive therapy
less intensive therapy
Survival
Time (day)

Figure 6

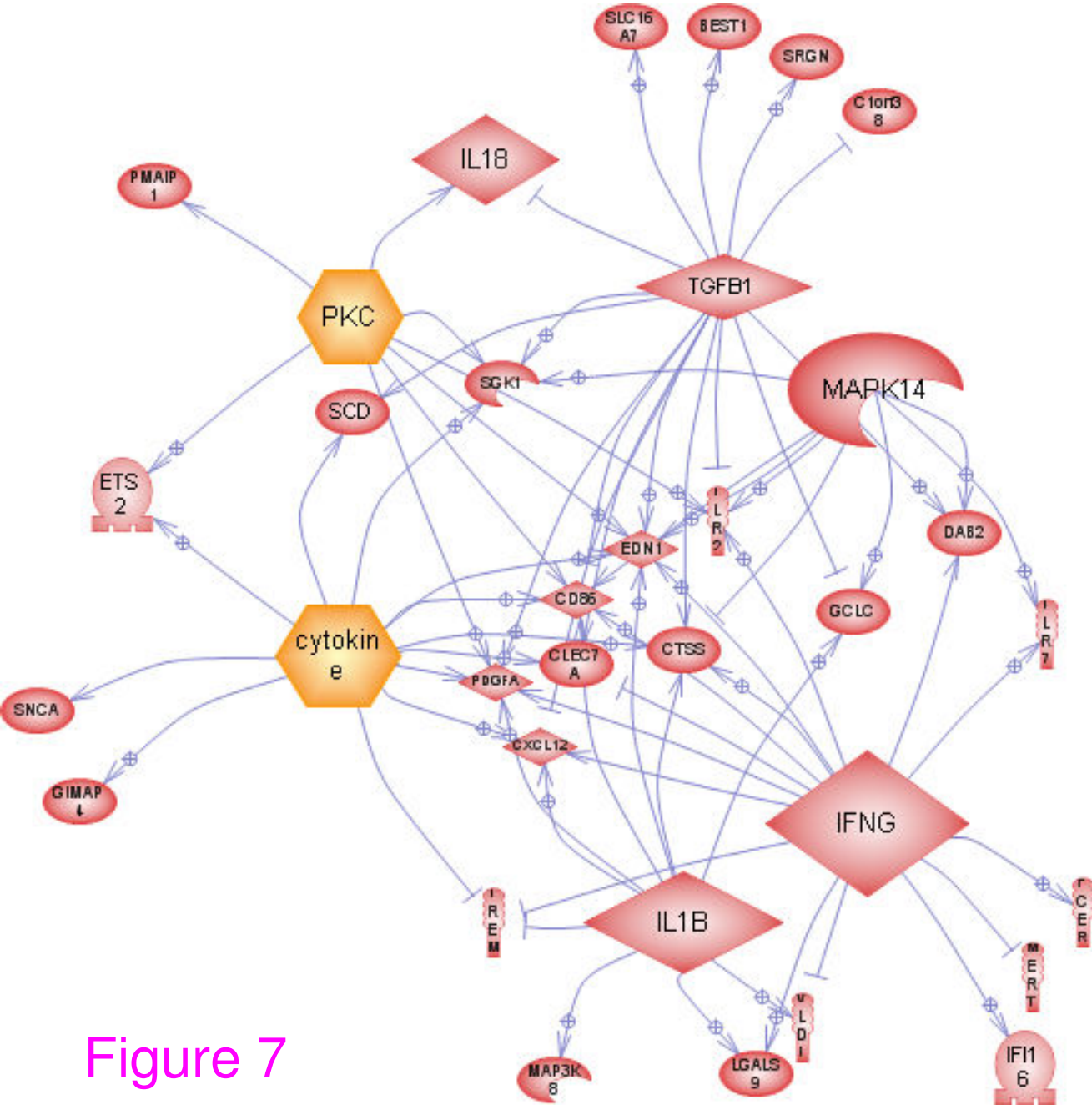

SLC16A7
BEST1
SRGN
C1orf38
IL18
PMAIP1
TGFB1
PKC
SGK1
MAPK14
SCD
ETS2
ILR2
DAB2
EDN1
CD86
GCLC
ILR7
cytokine
CLECTA
CTSS
PDGFA
SNCA
CXCL12
GIMAP4
IFNG
IREM
IL1B
MERT
VLDLI
MAP3K8
LGALS9
IFI16

Figure 7

Additional files provided with this submission:

Additional file 1: Additional file 1.pdf, 291K
http://www.biomedcentral.com/imedia/1679886300646454/supp1.pdf

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.