Published in final edited form as:

IEEE J Biomed Health Inform. 2014 May ; 18(3): 765–772. doi:10.1109/JBHI.2013.2276766.

Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis

Sonal Kothari,

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA

John H. Phan,

Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA and also with Emory University, Atlanta, GA 30332, USA

Todd H. Stokes,

Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA and also with Emory University, Atlanta, GA 30332, USA

Adeboye O. Osunkoya,

Department of Pathology, Emory University School of Medicine, Atlanta, GA 30322 USA

Andrew N. Young, and

Pathology and Laboratory Medicine, Emory University and Grady Health System, Atlanta, GA 30322 USA

May D. Wang [Senior Member, IEEE]

Department of Biomedical Engineering, School of Electrical and Computer Engineering, Winship Cancer Institute, Parker H. Petit Institute of Bioengineering and Biosciences, Institute of People and Technology, Georgia Institute of Technology Atlanta, GA 30322 USA and also with Emory University, Atlanta, GA 30332 USA

Sonal Kothari: sk9@gatech.edu; John H. Phan: jhphan@gatech.edu; Todd H. Stokes: todd.stokes@bme.gatech.edu; Adeboye O. Osunkoya: adeboye.osunkoya@emory.edu; Andrew N. Young: anyoung@gmh.edu; May D. Wang: maywang@bme.gatech.edu

SECTION I. Introduction

Computer-Aided decision support systems can guide pathologists in diagnosing cancer by extracting and modeling quantitative properties of histopathological images [1]. However, when histopathological images are acquired in different experimental setups and tested on pretrained diagnostic models, the prediction performance can suffer due to batch effects, i.e., nonbiological experimental variations such as age of sample, method of slide preparation, specifications of the imaging device, and type of postprocessing software [2]. Batch effects may lead to large differences in quantitative image features. Thus, it is difficult to accurately diagnose patients using prediction models trained with a separate batch. Because of batch effects, a pathology lab that uses multiple imaging devices (e.g., microscopes with mounted digital cameras or whole-slide scanners) may need to maintain multiple diagnostic models. Moreover, data acquired using older devices or experimental setups cannot be used in training models for future data acquired with newer devices/setups. This poses a huge

challenge for the cross-laboratory adoption and the standardization of pathology decision support systems.

Batch effects are also a major challenge for other biomedical data modalities. Although the causes of batch effects are different for each data modality, methods developed to remove batch effects may be applicable to multiple data modalities. For example, the sources of batch effects in microarray gene expression data include platform, laboratory, technician, and atmospheric ozone level [3], [4]. Batch effects generally affect the mean (location) and variance (scale/spread) of the data [5]. Therefore, batch effect removal methods focus on normalization of location and scale, e.g., ratio-based methods and ComBat [4], [5]. Luo et al. compared several batch effect removal methods for microarray data and found that ratio-based methods performed the best [4]. In a separate study, Chen et al. compared six batch effect removal methods and found that ComBat performs the best [3].

Removal of batch effects in histopathological images is a relatively new area of research [2]. However, with the emergence of large image data repositories such as the Cancer Genome Atlas, batch effects have become an increasingly important area of research. Histopathological image analysis studies have primarily focused on single-batch data [6]–[7] [8][9]. Some studies have highlighted color batch effects in histopathological images and suggested color normalization methods [10]–[11][12]. Color batch effects, which lead to variation in stain colors across batches, affect the performance of color segmentation methods and color features. Kothari et al. studied scale batch effects in histopathological images and suggested a scale normalization method based on nuclear area [13]. To the best of our knowledge, no published work quantifies or compares batch effect removal methods for histopathological images.

We compare six normalization methods including one image (scale) normalization method and five feature normalization methods: mean, rank, ratio, ComBatP, and ComBatN. Using four renal tumor histopathological datasets that are acquired using different experimental setups, we assess the impact of each batch effect removal method on image-based features and downstream prediction of renal tumor subtype and grade. Results indicate that data batch can be a larger source of variance in image features compared to biological factors such as grade and subtype. Most batch effect removal methods can reduce this variance to nearly zero. Moreover, batch effect removal methods can increase cross-batch and combined-batch prediction performance, with ComBatN performing the best.

SECTION II. Materials and Methods

A. Data

We use digital micrographs of renal tumor biopsy samples acquired in four experimental setups. Tissues samples in all four batches RCC1, RCC2, RCC3, and RCC4 were biopsied and fixed at Emory University. The micrographs for the first three batches were acquired at Emory University, while the fourth batch was acquired at the Georgia Institute of Technology. Table I lists image acquisition details for four batches. Each image is a rectangular section manually selected from a whole-slide image by a pathologist. Batches RCC1, RCC3, and RCC4 are annotated with both grade and subtype, while batch RCC2 is

annotated with only subtype. Images in these datasets represent one of three prominent renal tumor subtypes—chromophobe (CH), clear cell (CC), and papillary (PA)—and Fuhrman grade of one to four. Table II lists the number of samples per dataset for each subtype and grade.

Fig. 1 illustrates 512×512 -pixel subsections of three subtype samples in each of the four batches. These subtypes are histopathological subtypes and can be visually predicted based on morphology [14]. CC has clear cytoplasm with distinct cell membranes and round nuclei. CH has granular cytoplasm with prominent cell membranes, wrinkled nuclei, and perinuclear halos (i.e., white stain surrounding nuclei). PA has finger-like complex nuclear clusters. These properties are visually apparent in each of the four batches. However, images of each subtype appear very different between batches. In particular, RCC1 and RCC2 images appear to have different texture and scale compared to RCC3 and RCC4 images. Fig. 2 illustrates 512×512 -pixel subsections of four Fuhrman grade samples in three batches. G1 cells have small, intensely stained nuclei with no visible nucleoli. G2 cells have finely granular chromatin, slightly textured nuclei, and inconspicuous nucleoli. G3 cells have unequivocally recognizable nucleoli. G4 have nuclear pleomorphism (varying size of nuclei), hyperchromasia (abundance of DNA, leading to darker staining), and single to multiple macronucleoli. Similar to subtypes, grades are easy to distinguish within batches. However, grades across batches appear to be very different.

B. Image Feature Extraction Methods

The literature suggests several useful image features for predicting different cancer endpoints using histopathological images. Instead of extracting features reported specifically for renal tumor prediction, we adopt a generalized comprehensive feature set capturing different image properties such as color, texture, shape, and topology [15]. In total, we represent each image using 2663 image features (see Table III) and apply data mining approaches (discussed in Section II-D) to select optimal features and classification parameters. Since images in different batches are of different sizes (see Table I), we normalize the features that are affected by image size such as color or intensity histograms and nuclear count. Before extracting color features, we normalize image colors using colormap normalization [10].

Automatic feature extraction from histopathological images is complicated because of various levels of image segmentation. In our system, we use two levels of segmentation: color segmentation to separate different stains in the image, and nuclear segmentation to segment individual nuclei. Images used in this study are stained with Hematoxylin and Eosin (H&E), where nuclear, cytoplasmic, and glandular structures appear blue-purple, pink, and white, respectively. In addition, H&E-stained images may contain some intensely red structures that represent red blood cells. Pathologists easily distinguish between these structures by visual inspection. However, in the presence of color batch effects, these stains might appear very different [see Fig. 3(a)–(d)]. Therefore, it is a challenge for a computer to accurately segment these stains based on knowledge (supervision) from one batch [10]. To overcome this challenge, we adopt a three-step segmentation method that includes 1) color normalization to a set of standard reference images, 2) supervised segmentation and vote-

based labeling of pixels, and 3) segmentation refinement in the original color space [10]. We use a set of H&E-stained ovarian tumor images from The Cancer Genome Atlas [16]. The middle row of Fig. 3 includes examples of segmented stains from different batches. After color segmentation, we obtain a binary mask for nuclear stain. However, it is still difficult to isolate individual nuclei because: 1) texture in a high-grade nucleus can break it into segments [see Figs. 3(g)] and 2) adjacent nuclei can overlap, forming nuclear clusters [see Figs. 3(e) and 3(f)]. The first challenge can be addressed by merging neighboring nuclear stain structures. However, this may lead to even more complex clusters in images with large clusters, such as the papillary subtype. Thus, we only perform a merging treatment in images with a large percent of small isolated nuclear regions, i.e., images whose nuclear mask has greater than 10% percent of regions with area less than 20 pixels. For merging, we selectively grow the regions based on their area such that larger regions are grown more than smaller regions (which may represent noise). We grow regions morphologically using radial structural element with radius, $r=A/100$ pixels, where A is the area of a region. We limit the value of r to be 3 pixels. Based on empirical analysis, we found that this treatment works better than using a morphological closing operation on all regions. Next, we segment nuclear clusters using concavity detection and ellipse fitting [17]. The bottom row of Fig. 3 includes example images of segmented nuclei from different batches.

C. Normalization Methods

1. Scale Normalization—Upon visual inspection, we found that, besides color batch effects, which are handled using color normalization, these images differ in scale, i.e., images in batches RCC3 and RCC4 are at a higher scale compared to batches RCC1 and RCC2. Scale differences can be calculated using the physical size of a pixel. In digital micrographs, the physical pixel size varies with factors such as microscope magnification, CCD-pixel size, and digitizing software settings. The physical pixel size is not available for RCC1, RCC2, and RCC3 datasets. Therefore, we use a model based on nuclear area to estimate scale differences and normalize all batches [13]. Although nuclear area varies with subtype and grade in a batch, when we studied the distribution of all nuclei in a batch, we found that the distribution peaks at a specific nuclear area. Moreover, as we upscale or downscale the images, the distribution shifts right or left. Fig. 4(a) shows the distributions for nuclear area as RCC1 is scaled using the Lanczos (three-lobe) filter [13]. These distributions represent all nuclei in all images in RCC1. The scaling factor, s , affects both the x - and y -dimension and scales the nuclear area by a factor of s^2 . Therefore, the relationship between scaling factor, s , nuclear area in a scaled batch, A_s , and nuclear area with no

scaling, A_1 , is given by $s = \sqrt{\frac{A_s}{A_1}}$. We use median area of all nuclei in the batch at scale s to quantify A_s . In Fig. 4(b), we compare the empirical median nuclear area A_s of the RCC1 batch (red circles) when images are scaled for $s=0.5$ to 2, to the values predicted by the model using A_1 (cyan line). We can observe that values predicted by the model closely correspond to the empirical values. In Fig. 4(c), we have plotted the nuclear area distribution for four batches with no scaling. These plots suggest that the four batches are at different scales and RCC1 is at the lowest scale. Therefore, we scale down RCC2, RCC3, and RCC4 with scaling factors of 0.88, 0.53, and 0.46 [calculated using

$s = \sqrt{A_1(\text{new batch})/A_1(\text{RCC}_1)}$], respectively. In Fig. 4(d), we have plotted nuclear area distribution for four batches after scale normalization. After normalization, the distributions of the four batches are more similar. We then use scaled images to extract image features for prediction.

2. Mean Normalization—Scale normalization normalizes images to remove batch effects while all other removal methods (in this study), including mean, normalize features to overcome batch effects. After feature extraction, each image i in batch b is represented as a 2663-dimensional vector $Y_{i,b}$. The most prominent type of batch effect in microarray gene expression data results in differences in location (mean) and scale (spread) of gene expression values within batches. To overcome this challenge, researchers standardize the microarray data gene wise, where for every batch, expression of a gene over all samples is adjusted such that the mean is zero and the standard deviation is one. Similarly, in mean normalization, we normalize every feature f in batch b to be within a range $[0,1]$ using the

following transformation [18]: $\hat{Y}_f^{i,b} = \frac{1}{2} \left[\frac{(Y_f^{i,b} - \mu_f^b)}{3\sigma_f^b} + 1 \right]$, where μ_f^b and σ_f^b are the sample mean and standard deviation of feature f over samples in batch b , given by

$$\mu_f^b = \frac{1}{N_b} \sum_{i=0}^{N_b} Y_f^{i,b} \text{ and } \sigma_f^b = \sqrt{\frac{1}{N_b - 1} \sum_{i=0}^{N_b} (Y_f^{i,b} - \mu_f^b)^2}, \text{ where } N_b \text{ is the number of images in batch } b.$$

3. Rank Normalization—Unlike mean normalization, the rank normalization method does not assume an intrinsic normal distribution of features. Rank normalization also normalizes every feature j for a batch b to a range of $[0,1]$ [18]. We develop a rank function R_b , which orders the feature values $Y_f^{1,b}, Y_f^{2,b}, \dots, Y_f^{N_b,b}$ such that $R_b(Y_f^{m,b}) > R_b(Y_f^{n,b})$ implies $Y_m^{b,f} > Y_n^{b,f}$. Using this rank function, we normalize the features as follows:

$\hat{Y}_f^{i,b} = \frac{R_b(Y_f^{i,b}) - 1}{N_b - 1}$. To ensure that features are uniformly distributed in the range $[0,1]$ after rank normalization, we assign an average rank to images with the same feature value. For example, if there are n images with the same feature value, which is greater than the feature values of m images in a batch, we assign rank $m+n/2$ to all n images.

4. Ratio Normalization—Researchers have illustrated the usefulness of ratio-based methods for normalizing microarray expression data [4], where features are divided by the mean expression of a reference set of control samples (corresponding to each batch). In the absence of a reference set for a batch, researchers have used the mean expression of a batch in place of the control samples. However, this leads to information leakage because part of the labeled test samples is used for normalization as well. In our study, we have no reference set. Therefore, we normalize a feature f for a batch b using the median feature value for the

batch M_b^f , given by $\hat{Y}_f^{i,b} = \frac{Y_f^{i,b}}{M_b^f}$.

5. ComBat Normalization—Mean normalization adjusts location and scale for each feature independently. However, we can assume that batch effects are similar for most features and model batch effect-induced variance using accumulated knowledge across multiple features. Johnson et al. developed a Bayesian framework for modeling additive, γ_{bf} , and multiplicative, δ_{bf} , batch effects in microarray gene expression datasets [5]. They modeled batch effects using parametric (ComBatP) and nonparametric (ComBatN) priors. Similarly, we assume that $Y_{f,i,b}$, for feature f and image i in batch b , can be represented

using a location and scale (L/S) model, given by $Y_{f,i,b} = \alpha_f + X\beta_f + \gamma_f^b + \delta_f^b \varepsilon_f^{i,b}$, where α_f is the overall feature, X is a design matrix for sample conditions, β_f are regression coefficients corresponding to X , and ε is normally distributed error with mean zero and variance σ^2_f [5]. Johnson et al. illustrate that if the L/S model parameters are estimated using a Bayesian framework (which pools the information across features), parameters are more robust in removing batch effects. Therefore, we estimate $\alpha^b_f, \beta^b_f, \gamma^b_f$, and δ^b_f using the following suggested three-step Bayes framework: 1) calculate standardized data $Z_{f,i,b}$, which has similar overall mean and variance for all features, 2) estimate γ_{bf} and δ_{bf} using parametric (ComBatP) or nonparametric (ComBatN) priors, and 3) adjust the data for batch effects

using the following equation: $\hat{Y}_{f,i,b} = \frac{\hat{\sigma}_f}{\hat{\gamma}_f^b} (Z_{f,i,b} - \hat{\gamma}_f^b) + \hat{\alpha}_f + X\hat{\beta}_f$. We normalize datasets using the R language implementation of ComBat provided by Johnson et al. [5].

D. Feature Selection and Classification

We develop image-based prediction models for diagnosing renal tumor subtype and grade using data mining methods: feature selection and classification. Using the previously described methods, we represent images as a large set of normalized or unnormalized quantitative image features. However, only a few features among these features are informative for cancer prediction. Therefore, we select features using a supervised feature selection method, minimum-redundancy, maximum-relevance-difference (mRMR-D) [19]. We experiment with feature sizes in a range of 1 to 45. For classification, we develop prediction models using multiclass support vector machines (SVM) with the LIBSVM library, which returns the maximum voted label based on binary models [20]. We choose parameters for the SVM linear and radial basis kernels from the following list of cost values: $c = \{ \{0.1, 0.2, \dots, 0.9\}, \{1, 2, \dots, 9\}, \{10, 20, \dots, 100\} \}$. Also, for the radial basis kernel, we select gamma from: $\gamma = \{22, 21, \dots, 2-10\}$. We optimize all parameters—feature size, SVM cost, and SVM kernel—using ten iterations of three-fold cross-validation on the train set.

SECTION III. Results and Discussion

A. Variance in Data Contributed by Batch Effects

Batch effects often influence quantitative image features to such an extent that they are a major source of variation in the data. Thus, they often overwhelm the natural segregation of images due to biological classes. To illustrate the batch effect in the data under study, we clustered image features without any normalization from all samples in four batches [5]. Fig. 5 shows hierarchal clusters in the data, where the heat map highlights the feature (rows) variation across samples (columns). It can be observed that the four dominant clusters

correspond to four batches. We use principal variation component analysis (PVCA) to measure the variation in data contributed by the following factors: batch, grade, subtype, interaction between batch and subtype, interaction between batch and grade, and interaction between subtype and grade [3], [21]. PVCA is a useful method for calculating the proportion of variance attributable to different factors in high dimensional data. It is a combination of two popular data analysis methods: principal component analysis and variance component analysis (VCA). Before applying PVCA, we standardize the combined data (including three batches: RCC1, RCC3, and RCC4) using the same formula as mean normalization. We have excluded RCC2 in this study because it was not annotated with grades. The first step of PVCA involves reducing the dimensionality of data from 2663 features to the top few principal components (PCs) capturing a fixed portion (here 90%) of the variation (information) in the data. The second step involves applying VCA to calculate the variation in each PC contributed by each factor. For calculating the variance, VCA assumes each factor as a random effect in a linear mixed model. Variances for each PC are weighted by Eigenvalues of the component and averaged. Fig. 6 illustrates weighted variances contributed by each factor in features with and without normalization. Variance contributed by batch is 0.365 in features without normalization, which is much higher than biological factors: grade and subtype. Variance contributed by batch is considerably reduced by all feature normalization methods: mean, rank, ratio, ComBatP, and ComBatN.

We visualize the distribution of samples in the image feature space using scatter plots of component scores for the first and second PCs [4]. Fig. 7 illustrates the scatter plots with and without normalization. The scatter plots show clear separation of samples from different batches (represented by four colors) in the data without normalization and scale normalization. In contrast, samples are randomly distributed in the plots for all other normalization methods.

To further investigate the impact of batch effects on image features in scale-normalized data, we ranked image features that are predictive of batch using mRMR feature selection. Comparing the ranked list for scale normalized data to unnormalized data, we found the following: 1) texture properties, which are highly predictive of batches, are ranked high in both datasets and 2) shape-based properties such as median boundary fractal and Fourier error are informative for batch prediction in unnormalized data but not in scale normalized data. Therefore, besides scale, batches can differ in texture, which is not corrected by scale normalization. Differences in image formats associated with different compression methods, camera CCD, and magnification are possible causes for texture differences in the batches.

B. Within-Batch Prediction Establishes Baseline Performance

To verify the utility of the proposed image feature set for classifying renal grades and subtypes, we perform within-batch cross-validation of prediction models for all batches. We observe that during cross-validation within a batch, prediction models perform very well with classification accuracy greater than 88% for all cases except RCC1 grading (see Table IV). A possible cause for the low accuracy of the grading model for RCC1 could be the small sample size, i.e., only 53 total samples are available in four classes.

C. Batch Effect Removal Methods Increase Cross-Batch Prediction Performance

We perform a cross-batch validation of grading and subtyping models with and without normalization. In total, with all combinations of train set, test set, and endpoint, we have 18 comparisons. Fig. 8 illustrates performance accuracies for all comparisons and normalization methods. Entries in the figure are highlighted in pink or blue if the performance has increased or decreased compared to no normalization. In general, the performance of prediction models is much lower compared to cross-validation within a batch. Possible reasons for this decrease are as follows: 1) biological variance in grade or subtype is not sufficiently captured by a train batch, or 2) normalization methods are not able to completely eliminate the batch effects. However, normalization methods do improve the prediction performance of several models. When compared to prediction accuracy of features with no normalization, we observed that the ComBatN, ComBatP, mean, and rank normalization methods resulted in average performance increases of 16%, 14%, 14%, and 12%, respectively. Moreover, ComBatN resulted in the largest number of cases with prediction improvement, with 83%.

D. Performance of Prediction Models After Batch Combination

In a clinical setting, a decision support system is often trained with a set of images collected in batches over time. In such a scenario, batch removal methods are essential before combining the batches. We compare the prediction performance of renal endpoints with and without normalization while combining two or more batches for training (see Fig. 9). Similar to cross-batch validation, all normalization methods except scale normalization significantly improve the performance compared to no normalization. We observed that the mean, ComBatN, rank, ComBatP, and ratio normalization resulted in average performance increases of 15%, 14%, 14%, 13%, and 11%, respectively. Moreover, ComBatN resulted in the largest number of cases with prediction improvement, with 90%. With these results, we can conclude that, even with the presence of data from multiple batches in the train set, feature selection cannot filter out features that are significantly affected by batch effects. As such, it is essential that the features are normalized before batch combination. Two prediction models in Fig. 9 combine all batches. Thus, we report cross-validation performance in these cases. The cross-validation performances are very high and comparable to within-batch prediction performances. Thus, representation of test set samples by including similar samples (i.e., same batch) in the train set can significantly improve the performance.

E. Effect of Normalization Methods on Image Integrity

Normalization methods can introduce another level of image processing in the decision making system. This processing can affect image properties and features. For instance, we use the Lanczos filter for down-sampling images in scale normalization, which may cause aliasing artifacts in the scaled images and affect texture features. Unlike scale normalization, feature normalization is performed after feature extraction and does not affect image integrity. This may be one reason for the poor performance of the scale normalization method.

SECTION IV. Conclusion

We have investigated batch effects in histopathological images and compared multiple batch effect removal methods. Although the presence of batch effects is an important challenge for computer-aided decision support systems, only a few researchers have investigated the impact of batch effect removal methods on histopathological image classification. Using four renal tumor image batches that have been annotated with cancer grade and subtype, we found that, compared to no normalization, ComBatN, ComBatP, mean, rank, and ratio normalization methods improve the performance of predicting tumor grade and subtype. In particular, ComBatN performs the best in terms of average increase in the performance and total number of prediction cases with an increase in the performance. Investigation of batch effects in histopathological images may become increasingly important as data repositories expand to contain valuable clinical knowledge from multiple institutions.

REFERENCES

1. Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.* 2009 Oct.2:147–171. [PubMed: 20671804]
2. Phan JH, Quo CF, Cheng C, Wang MD. Multiscale integration of -omic imaging and clinical data in biomedical informatics. *IEEE Rev. Biomed. Eng.* 2012 Dec.5:74–87. [PubMed: 23231990]
3. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE.* 2011; 6:e17238. [PubMed: 21386892]
4. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H, Zhao C, Elloumi F, Shi W, Thomas R, Lin S, Tillinghast G, Liu G, Zhou Y, Herman D, Li Y, Deng Y, Fang H, Bushel P, Woods M, Zhang J. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogeno. J.* 2010; 10:278–291.
5. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007; 8:118–127. [PubMed: 16632515]
6. Po-Whei H, Cheng-Hsiung L. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Trans. Med. Imag.* 2009 Jul; 28(7):1037–1050.
7. Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. *IEEE Trans. Biomed. Eng.* 2003 Jun; 50(6):697–704. [PubMed: 12814236]
8. Basavanahally AN, Ganesan S, Agner S, Monaco JP, Feldman MD, Tomaszewski JE, Bhanot G, Madabhushi A. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.* 2010 Mar; 57(3):642–653. [PubMed: 19884074]
9. Sertel O, Kong J, Catalyurek U, Lozanski G, Saltz J, Gurcan M. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. *J. Signal Process. Syst.* 2009; 55:169–183.
10. Kothari S, Phan JH, Moffitt RA, Stokes TH, Hassberger SE, Chaudry Q, Young AN, Wang MD. Automatic batch-invariant color segmentation of histological cancer images. *Proc. IEEE Int. Symp. Biomed. Imag.* :657–660.
11. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun G, Schmitt C, Thomas NE. A method for normalizing histology slides for quantitative analysis. *Proc. IEEE Int. Symp. Biomed. Imag.* :1107–1110.
12. Magee D, Treanor D, Crellin D, Shires M, Smith K, Mohee K, Quirke P. Colour normalisation in digital histopathology images. *Proc. Opt Tissue Image anal. Microscopy Histopath. Endoscopy (MICCAI Workshop).* :100–111.
13. Kothari, S.; Phan, JH.; Wang, MD. Scale normalization of histopathological images for batch invariant cancer diagnostic models; *Proc. IEEE Eng. Med. Biol. Soc. Conf.*; p. 4406-4409.

14. Eble, J.; Sauter, G.; Epstein, J.; Sesterhenn, I. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. IARC press; 2004.
15. Kothari, S.; Phan, JH.; Young, AN.; Wang, MD. Histological image feature mining reveals emergent diagnostic properties for renal cancer; Proc. IEEE Int. Conf. Bioinformat. Biomed; p. 422-425.
16. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
17. Kothari, S.; Chaudry, Q.; Wang, MD. Extraction of informative cell features by segmentation of densely clustered tissue images; Proc. IEEE Eng. Med. Biol. Soc. Conf; p. 6706-6709.
18. Aksoy S, Robert M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recog. Lett. 2001; 22:563–582.
19. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J. Bioinformat. Comput. Biol. 2005; 3:185–205.
20. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011; 2:27.
21. Boedigheimer M, Wolfinger R, Bass M, Bushel P, Chou J, Cooper M, Corton JC, Fostel J, Hester S, Lee J, Liu F, Liu J, Qian H, Quackenbush J, Pettit S, Thompson K. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics. 2008; 9:285. [PubMed: 18549499]

Fig. 1.

Image samples of three subtypes in four batches of renal cell carcinoma. Each image is a 512×512 -pixel subsection of the original whole-slide sample.

Fig. 2.

Image samples of four renal cell carcinoma grades in three batches. Each image is a 512×512 -pixel subsection of the original whole-slide sample.

Fig. 3.

Segmentation results for sample renal cell carcinoma histopathological images from four data batches illustrate scale batch effects. (a–d) 512×512 subsections of images from RCC1, RCC2, RCC3, and RCC4, respectively. (e and f) Pseudocolored color segmentation results, where blue, pink, white, and red represent pixels in nuclei, cytoplasm, glands, and red blood cells, respectively. (i–l) Nuclear segmentation results, where nuclei are marked using green ellipses on the original images (a–d). Note the difference in scale and nuclear size among four batches.

Fig. 4.

Scale normalization of images based on median nuclear area model. (a) Nuclear area distributions of RCC1 at different scales, (b) comparison between empirical and model-based nuclear area values for RCC1, (c) nuclear area distributions for four batches without scaling, and (d) nuclear area distributions for four batches after scale normalization. Scaling factors of other three batches with respect to RCC1 are marked in B.

Fig. 5.

Unsupervised hierarchical clustering of renal cell carcinoma histopathological image features in four datasets illustrates batch effects. Columns in the heatmap correspond to samples while rows correspond to individual features. Heatmap is normalized to show variation of a feature across different samples, where values above, at, and below the mean are represented as red, black, and green colors, respectively. The dendrogram and horizontal colored bars above the heatmap illustrate clusters in the data. It can be observed that samples are primarily clustered based on their batch. Note: clusters without color bars contains samples from multiple batches.

Fig. 6.

PCVA on a combined dataset, including RCC1, RCC3, and RCC4, with and without normalization. Bars indicate weighted variance contributed by various factors to the overall variation. Variance contributed by batch is significantly reduced by feature normalization methods.

Fig. 7.

Scatter plots representing scores of samples for first and second PCs. The percentage in the bracket represents percent of total variation (energy) along that component. With no normalization, batches (represented by four colors) are clearly separated. With all normalization methods except scale, separation between batches is reduced and samples seem to be randomly distributed.

Fig. 8.

Cross-batch validation accuracy of renal prediction models. The performance of normalized models is highlighted based on the change in prediction accuracy compared to no normalization. Feature normalization methods, especially ComBatN, increase the prediction performance for most of the models.

Fig. 9.

Accuracy of renal prediction models after batch combination. The performance of normalized models is highlighted based on the change in prediction accuracy compared to no normalization. Feature normalization methods, especially ComBatN, increase the prediction performance for most models. The performance significantly increases with representation of multiple batches in the train set.

Table 1

Factors	RCC1	RCC2	RCC3	RCC4
Microscope	Nikon Eclipse 80i	Nikon Eclipse 80i	Olympus BX51	Zeiss Axio Imager z2
Magnification	20×	20×	40×	40×
Camera	Nikon DS-2MV	Nikon DS-2MV	Olympus DP71	Zeiss AxioCam MRm
CCD pixel size (μm)	4.40 × 4.40	4.40 × 4.40	4.40 × 4.40	6.45 × 6.45
Year	2008	2006	2011	2012
Subjects	15	12	12	18
Images	53	36	72	160
Image Format	JPEG	JPEG	PNG	PNG
Image size (Pixels)	1600×1200	1600×1200	2040×1536	2040×1536

Table 2

		RCC1	RCC2	RCC3	RCC4
Subtype	CH	20	12	26	30
	CC	17	12	24	82
	PA	16	12	22	48
Grade	G1	13	N/A	18	20
	G2	13	N/A	18	63
	G3	13	N/A	18	50
	G4	14	N/A	18	27

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Type	Description	Count
Color	R, G, and B histograms	48
Texture	Haralick, Gabor, wavelet packet, multiwavelet of a 64-level color quantized image; <i>Haralick, Gabor, wavelet packet, multiwavelet</i> , and gray-level distribution of a grayscale image; Haralick, Gabor, wavelet packet, multiwavelet, and stain co-occurrence of four-level grayscale image; and Haralick and gray-level distribution of graylevel images of nuclear and cytoplasmic regions	1707
Shape	count, pixel area, convex hull area, solidity, perimeter, elliptical properties (area, major-minor axes lengths, eccentricity and orientation), boundary fractal, bending energy and Fourier reconstruction error of nuclear, cytoplasmic, and glandular shapes; elliptical properties and count of segmented nuclei	796
Topology	Delaunay triangle (area and side length), Voronoi diagram (area, side length and perimeter), minimum spanning tree edge length and closeness of spatial graphs using centers nuclear shapes and individual segmented nuclei	112

Table 4

	RCC1	RCC2	RCC3	RCC4
Subtype	0.88	0.88	0.93	0.91
Grade	0.44	NA	0.99	0.88

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript