# Automatic color unmixing of IHC stained whole slide images

D.J. Geijs[a, b], M. Intezar[a], J.A.W.M. van der Laak[a], and G.J.S. Litjens[a]

[a]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[b]University of Twente, Enschede, The Netherlands

## ABSTRACT

Assessment of immunohistochemically stained slides is often a crucial diagnostic step in clinical practice. However, as this assessment is generally performed visually by pathologists it can suffer from significant inter-observer variability. The introduction of whole slide scanners facilitates automated analysis of immunohistochemical slides. Color deconvolution (CD) is one of the most popular first steps in quantifying stain density in histopathological images. However, color deconvolution requires stain color vectors for accurate unmixing. Often it is assumed that these stain vectors are static. In practice, however, they are influenced by many factors. This can cause inferior CD unmixing and thus typically results in poor quantification. Some automated methods exist for color stain vector estimation, but most depend on a significant amount of each stain to be present in the whole slide images. In this paper we propose a method for automatically finding stain color vectors and unmixing IHC stained whole slide images, even when some stains are sparsely expressed. We collected 16 tonsil slides and stained them for different periods of time with hematoxylin and a DAB-colored proliferation marker Ki67. RGB pixels of WSI images were converted to the hue saturation density (HSD) color domain and subsequently K-means clustering was used to separate stains and calculate the stain color vectors for each slide. Our results show that staining time affects the stain vectors and that calculating a unique stain vector for each slide results in better unmixing results than using a standard stain vector.

**Keywords:** histopathology, whole slide imaging, color deconvolution, HSD color space

## 1. INTRODUCTION

Microscopic assessment of histopathological slides is a crucial diagnostic step in clinical practice. In histopathology, immunohistochemical (IHC) stains are often used to provide contrast between non-relevant tissue and a target of interest (e.g. a specific protein). The introduction of digital slide scanners has led to an interest in applying image analysis techniques to provide reproducible quantitative parameters used as imaging biomarkers.

Color deconvolution (CD)[1] is one of the most popular methods to quantify the stain density in histopathological images. The technique allows to separate stains by exploiting differences in the light absorption spectra of differently colored stains. CD does require a priori knowledge of color vectors (RGB) of each specific stain. CD proves to be a valuable tool and is widely used in different applications such as automatic nuclei detection and tissue segmentation.[2]

CD algorithms produce good results for similar sets of images, but often struggle with multi-center data. The RGB color vectors needed for CD algorithms are influenced by many factors, e.g. the staining protocol, type of tissue fixation, the specific antibody and the slide digitization process.[3] All these factors can differ substantially between centers and even between different batches of slides from the same center. This requires defining the color vector for each slide manually, avoiding inferior CD unmixing results which can affect the whole algorithm pipeline. This essentially hampers the applicability of using such algorithms for big clinical datasets, since manual determination of the color vectors is tedious and time consuming.

In the same year Macenko *et al.*[4] and Magee *et al.*[5] proposed a method for estimating stain vectors. Magee *et al.* used a supervised pixel classification-based approach to estimate stain colors, whereas Macenko *et al.* used a singular value decomposition (SVD)-based approach to directly estimate the matrices. Niethammer *et al.*[6] extended the stain matrix estimation method of Macenko et al. to improve stability in cases where images contain uneven proportions of each stain.[7] These studies focused mainly on unmixing H&E stained slides where both hematoxylin and eosin content are high.[7]

Finding the color components in an IHC slide is a more challenging task because of scarceness of high intensity IHC stain color. In this research we propose an automated method for finding color vectors using k-means clustering in a hue saturation density (HSD) color space[8] and unmixing IHC stained whole slide images (WSI) . The algorithm is evaluated on tonsil tissue sections, which were immunohistochemically stained with Ki67 proliferation marker. Resulting slides display sparse positive areas. We also analyze the effect of staining time on the color vectors and compare final unmixing results with reference and standard color vectors.

## 2. METHODOLOGY

### 2.1 Data

From a single tonsil tissue block 18 slides were cut and stained with Ki67 IHC. Ki67 was visualized using 3,3'-diaminobenzidine (DAB) and slides were counterstained using hematoxylin. For each individual slide there was a single hematoxylin, single DAB and a combined hematoxylin/DAB-staining available. First all 18 slides were stained with a hematoxylin stain. After staining, the slides were scanned with the 3DHistech Pannoramic 250 Flash II whole slide image scanner. The slides were destained and treated with a Ki67-DAB staining and scanned. After scanning the slides, another hematoxylin staining procedure was performed. The staining time was varied for hematoxylin and Ki67 primary antibody incubation. Three different staining times for each stain were chosen resulting in nine combinations performed in duplo. Time periods for hematoxylin and Ki67 were 1, 30 and 60 seconds and 5, 10 and 15 minutes, respectively.

### 2.2 Automatic unmixing

A prerequisite of color deconvolution is a pre-defined stain matrix consisting of three stain vectors. The stain vector describes the proportion of each wavelength absorbed by the stain. Since RGB images are used, the method is restricted to those three wavelengths. The procedure described in this paper was developed to automatically determine the stain vector for two stains in a whole slide image. The stain matrix is used to either deconvolve the whole image or specific tiles for further analysis.

All the RGB color values discussed in this paper were converted to their optical density (OD),

$$OD = -\log_{10}(\frac{I}{I_0}) \tag{1}$$

where I is the RGB color vector with each component normalized to [0,1] by dividing with the background RGB color vector $I_0$. A default background of $I_0 = [255, 255, 255]$ was used and $I > 0$. The WSI was processed in a tiled-by-tile approach. Pixels of a tile were converted from OD to coordinates in the HSD space[8] , defined by $c_x, c_y$ :

$$c_x = \frac{OD_r}{OD_{rgb}} - 1 \tag{2}$$

$$c_y = \frac{OD_g - OD_b}{OD_{rgb}\sqrt{3}} \tag{3}$$

where $OD_{rgb}$ is the average of the OD for each RGB color channel, calculated by Eq. 1 . The transformation resulted in a cloud of points that was projected in the HSD chromaticity triangle. Points with a low optical density ($OD < 0.3$) were removed to reduce noise.

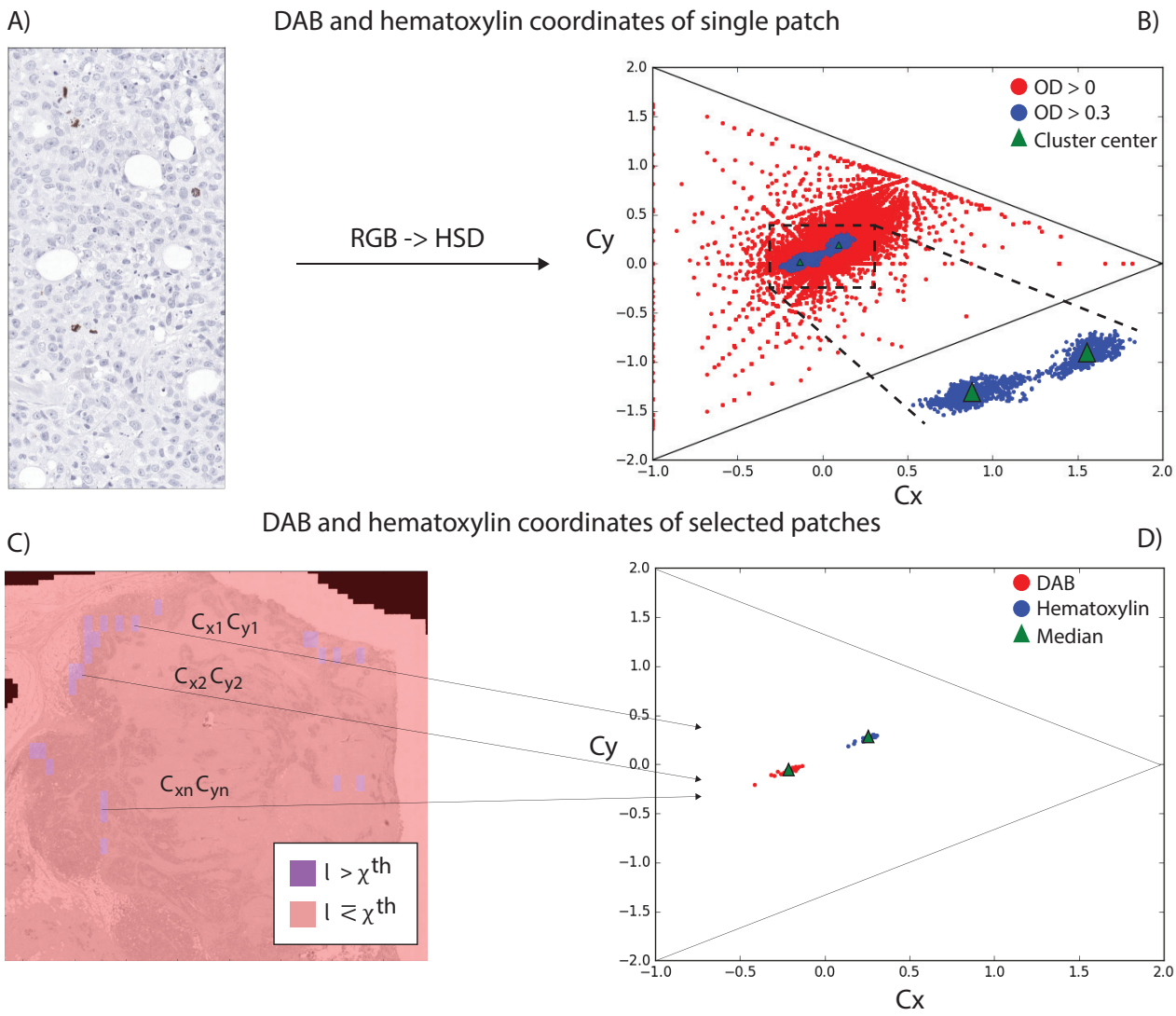


Figure 1. A) Example tile extracted from IHC WSI. The RGB values of the patch were converted to the HSD domain. B) Scatterplot of all pixels from the extracted patch. The boundaries of the HSD domain are plotted to indicate the range in which a pixel can be transformed. Red circles indicate the $c_x c_y$ position of pixels with $OD > 0$. Plotted in blue only the pixels with an $OD > 0.3$ and shows a huge reduction in noise. K-means clustering is used to find the cluster centers caused by the presence of the two stains. C) A part of a heatmap of a WSI, the purple squares correspond to tiles with an Euclidean distance $l > \chi^{th}$. The purple boxes level of detail and dimensions correspond to the example tile above (A). D) A visualization of all collected cluster centers of tiles with $l > \chi^{th}$. The median of this group is the final color vector used for deconvolution.

For each tile, K-means clustering was used on the $c_x c_y$ data to detect two clusters ($k = 2$) and to find their centers. The K-means algorithm starts by randomly choosing two cluster centers and adding each of the observations to the nearest of those clusters, it updates the cluster center and iterates until it converges to a final solution. These final cluster centers represented the two different stains that were present: hematoxylin and DAB. After detecting the two clusters the Euclidean distance was calculated between the two cluster centers. Even when there was only one stain present in the tile the K-means clustering was forced to detect two clusters. Centers close to each other (i.e. only one stain present) thus resulted in a much smaller Euclidean distance.

The tiles that measured the largest Euclidean distance between the two clusters were selected. These tiles

were regarded to be likely to have two different stains. This was done by calculating the $\chi^{th}$ percentile from all the distances. Tiles with distance $l > \chi^{th}$ were selected. $\chi = 99$ was used in this paper. The median of each cluster was selected as the $c_x c_y$ for either hematoxylin or DAB. The stain vector was calculated by reversing Eq. 2 and Eq. 3 with $OD_{rgb} = 1$.

### 2.3 Validation

The $c_x c_y$ coordinates of the reference single stain slides were measured in a manual ROI. For each stain condition an average reference $c_x c_y$ was calculated. Our automated method was tested on each corresponding double stained hematoxylin/DAB-slide. The distance between the average reference $c_x c_y$ and the automatic $c_x c_y$ was calculated and noted as $l_{auto}$. The same was done for the standard, static $c_x c_y$ and noted as $l_{std}$. The standard is the stain vector proposed by Ruifrok *et al.*[1] Last, the quality of our method was evaluated by comparing the reconstructed images after deconvolution qualitatively.

## 3. RESULTS

We tested the effect of different staining times on stain color vector distribution. Reference $c_x c_y$ values, manually calculated based on single stained slides, clearly show two clusters (Fig. 2) without overlap, representing to hematoxylin and DAB. Figure 2 shows the spread of the manually acquired points under different time conditions. The scatterplot containing the manually determined hematoxylin coordinates shows three different clusters based on the staining time, which is not observed in the DAB scatterplot.

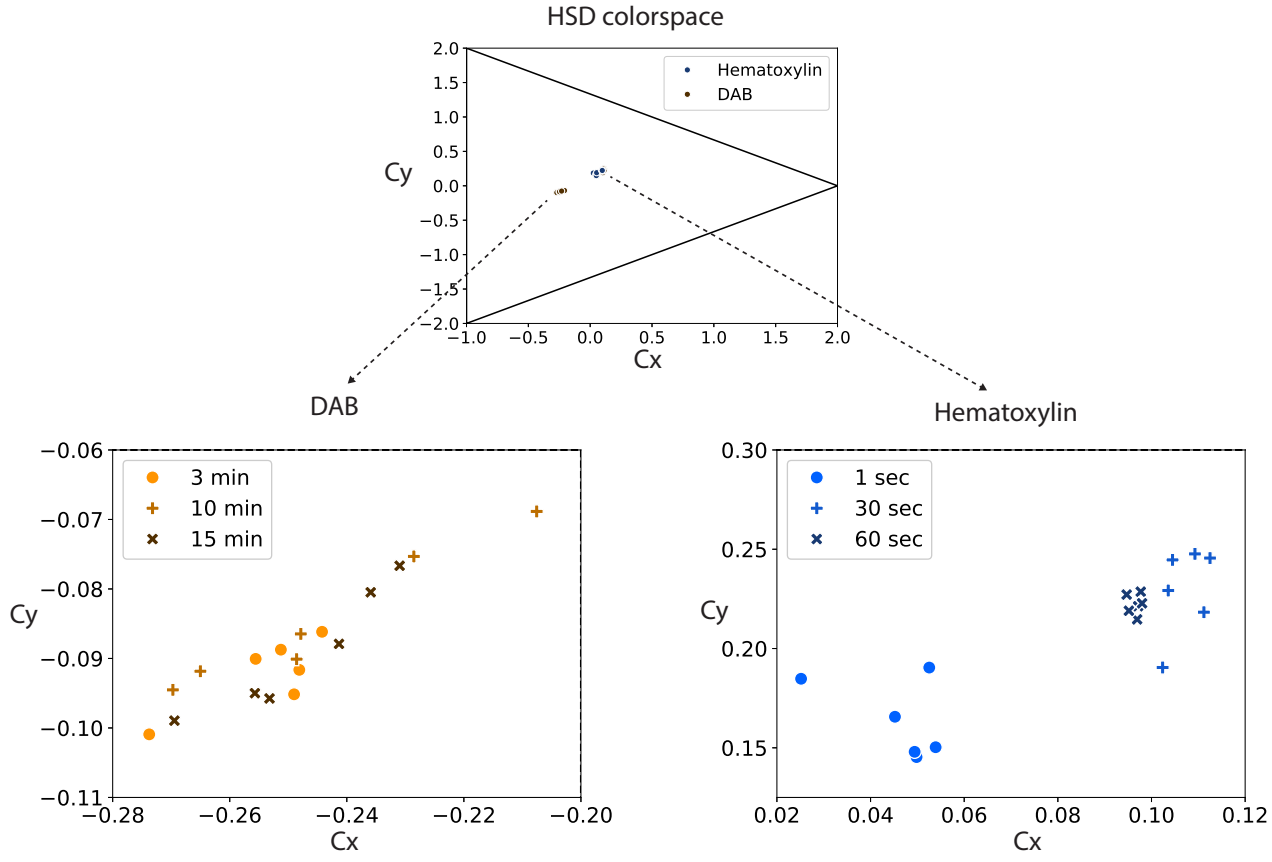


Figure 2. A HSD colorspace containing $c_x c_y$ coordinates of single stained slides. DAB and Hematoxylin coordinates are determined by selecting a manual patch and are used as reference. Both stains are treated for different time conditions.

Table 1. Distances of automatic ($l_{auto}$) and standard($l_{std}$) $c_x c_y$ coordinates compared to manually measured reference $c_x c_y$ coordinates. For each stain condition $1\sigma$ is calculated.

| Stain condition | Group size | $l_{auto} \pm \sigma$ | $l_{std}$ |
|---|---|---|---|
| Hematoxylin 1 s | 6 | 0.143 ± 0.064 | 0.668 |
| Hematoxylin 30 s | 6 | 0.067 ± 0.013 | 0.756 |
| Hematoxylin 60 s | 6 | 0.054 ± 0.004 | 0.743 |
| DAB 3 min | 6 | 0.149 ± 0.075 | 0.278 |
| DAB 10 min | 6 | 0.109 ± 0.029 | 0.290 |
| DAB 15 min | 6 | 0.105 ± 0.053 | 0.285 |

After manual determination of the stain vectors, the vectors are computed automatically applying the proposed algorithm and both are compared. The measurements are shown in Table 1. The algorithm was successful in finding both stains in all 18 slides. The Euclidean distance to the reference coordinate is measured and averaged for all coordinates belonging to each stain condition together with the standard deviation $\sigma$. The sigma is a measure for the spread and calculated for each stain condition. The slides treated with hematoxylin for 1 second show the largest $l_{auto}$ and $\sigma$, the same holds for slides treated with DAB for 3 minutes.

When the stain vector is calculated with the found DAB and hematoxylin coordinates the image can be deconvolved. In Figure 3 images are reconstructed using stain vectors found by the automatic method, found manually (standard) and the standard stain vector described by Ruifrok et al. The standard vector gives a brighter purple hematoxylin signal and lighter brown DAB signal. The automatic and manual methods show a bigger resemblance compared to the standard vector. The automatic and manual unmixed DAB channels have a darker brown color, with the automatic DAB channel almost being black.

## 4. DISCUSSION

The slides stained for 60 seconds with hematoxylin form dense clusters in the $c_x c_y$ space and show the least amount of heterogeneity compared to the 30 seconds and 1 second condition. Thus, introducing differences in staining time results in color differences and optical density values and the $c_x c_y$ values of the stain. The shortest staining time for DAB and hematoxylin both have the largest distance and spread. This may be explained by the fact that the stain has insufficient time to fully saturate, resulting in lower concentrations of light absorbing product. This in turn will results in a low optical density, with reduced signal to noise ratio and a larger variability.

The distances between the reference data and the output of the algorithm show that all found $c_x c_y$ coordinates have a distance $l_{auto} < l_{std}$, indicating that the proposed algorithm outperforms the use of static stain vectors. The reconstruction results in observable color differences between the original and unmixed images for all three methods. Comparable results can be observed for unmixing by the reference method and the newly proposed algorithm. In contrast, the use of predefined stain vectors visually appears to be less accurate. As can be observed, the DAB signal resulting from unmixing using the predefined vectors contains signal in locations containing DAB negative nuclei, which is undesirable. This phenomenon is not present in results from the proposed algorithm and only to a moderate degree for the reference method. The contrary, presence of hematoxylin in DAB positive nuclei is to be expected but can not be observed for the results from the automatic and reference method. This may be attributed to the fact that DAB chromogen is not a true light absorber. DAB exhibits light scattering behavior instead, which is not perfectly modeled by the Beer-Lambert law.

Overall, our results show that the use of predefined, fixed stain vectors is not advisable because of large variability between different IHC stained slides. The HSD space was found to facilitate extraction of stain vectors. Manually guided extraction and use of the newly proposed algorithm both result in reliable stain vectors, the latter being less labor intensive and potentially better reproducible.

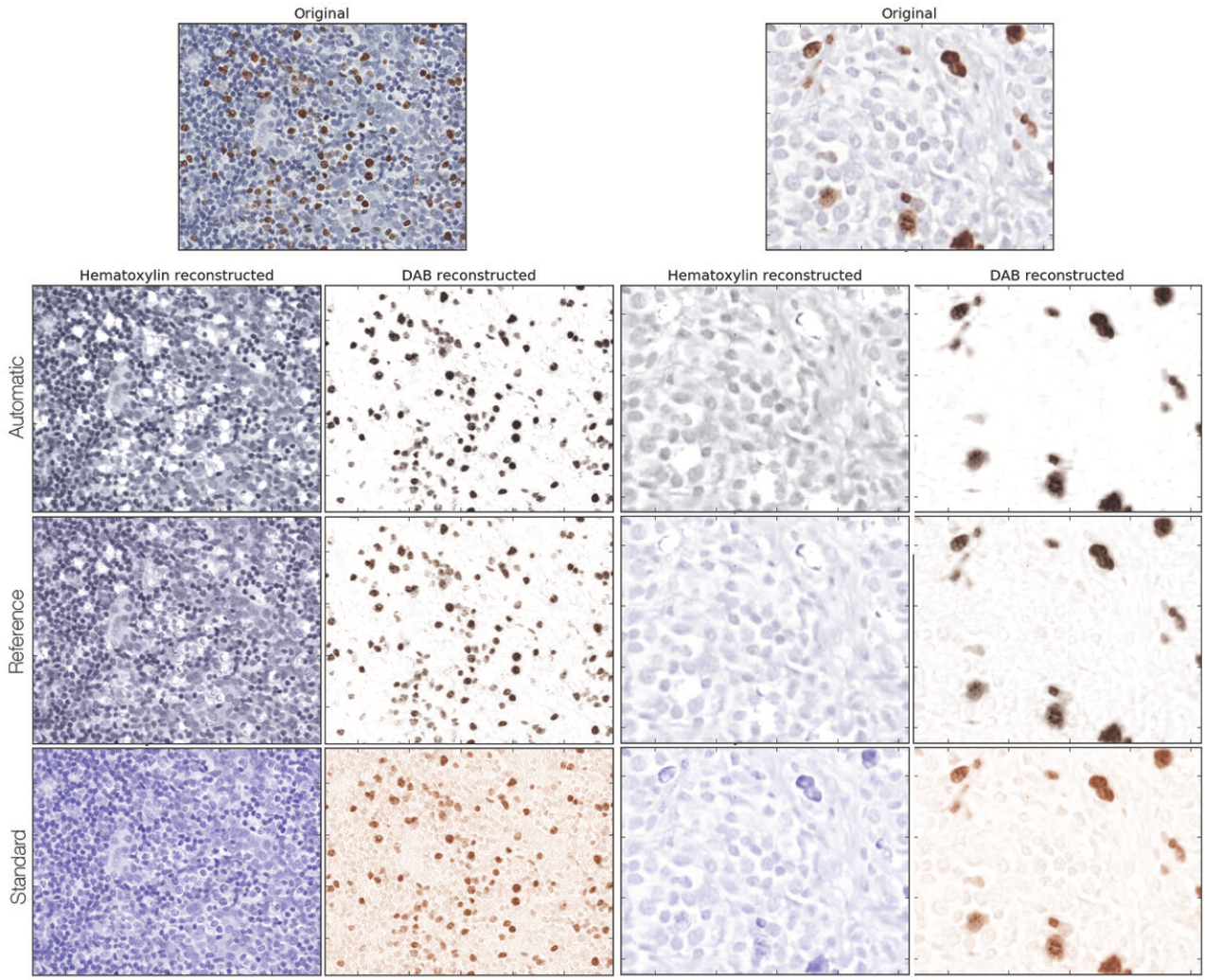


Figure 3. Image reconstruction compared after deconvolution using a stain vector calculated from: *automatic* $c_x c_y$ coordinates, manual *reference* $c_x c_y$ coordinates and *standard* $c_x c_y$ coordinates. left) Patch of a whole slide image stained with hematoxylin for 60 seconds and DAB 10 minutes. right) Patch of whole slide image stained with hematoxylin for 1 seconds and DAB 10 minutes. The colors used for visualizing the reconstructed images correlate to the automatic stain vectors found in the HSD domain. DAB reconstructed with *standard* and *reference* coordinates show positive stained nuclei which should be negative. This indicates inaccuracies in the stain vectors.

## REFERENCES

[1] Ruifrok, A. C., Johnston, D. A., et al., "Quantification of histochemical staining by color deconvolution," *Analytical and quantitative cytology and histology* **23**(4), 291–299 (2001).

[2] Irshad, H., Veillard, A., Roux, L., and Racoceanu, D., "Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential," *IEEE reviews in biomedical engineering* **7**, 97–114 (2014).

[3] Onder, D., Zengin, S., and Sarioglu, S., "A review on color normalization and color deconvolution methods in histopathology," *Applied Immunohistochemistry & Molecular Morphology* **22**(10), 713–719 (2014).

[4] Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," in [*Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*], 1107–1110, IEEE (2009).

[5] Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K., and Quirke, P., "Colour normalisation in digital histopathology images," in [*Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*], **100**, Daniel Elson (2009).

[6] Niethammer, M., Borland, D., Marron, J., Woosley, J., and Thomas, N., "Appearance normalization of histology slides.," in [*Machine learning in medical imaging. MLMI (Workshop)*], **6357**, 58–66, NIH Public Access (2010).

[7] Khan, A. M., Rajpoot, N., Treanor, D., and Magee, D., "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering* **61**(6), 1729–1738 (2014).

[8] van der Laak, J. A., Pahlplatz, M. M., Hanselaar, A. G., and de Wilde, P., "Hue-saturation-density (hsd) model for stain recognition in digital images from transmitted light microscopy," *Cytometry Part A* **39**(4), 275–284 (2000).