



# A New Paradigm of RNA-Signal Quantitation and Contextual Visualization for On-Slide Tissue Analysis

Auranuch Lorsakul<sup>1</sup>(✉) and William Day<sup>2</sup>(✉)

<sup>1</sup> Roche Tissue Diagnostics, Imaging and Algorithms, Digital Pathology, Santa Clara, CA, USA

auranuch.lorsakul@roche.com

<sup>2</sup> Roche Tissue Diagnostics, Tissue Research and Early Development, Tucson, AZ, USA

bill.day@roche.com

**Abstract.** An objective digital pathology solution to quantify the ribonucleic acid (RNA) signal in tissue samples could enable analysis of gene expression changes in individual cancer and dysregulated normal cells (immune cells, etc.). Here, we present a new method that leverages the punctate RNA *In-situ* hybridization (ISH) signal to quantify gene expression, while maintaining tissue context and enabling single cell analysis and workflow. This digital pathology solution detects and quantifies the punctate dot signals generated by one- and two-color RNA ISH technology in formaldehyde fixed-paraffin embedded (FFPE) tissue. The digital pathology solution was implemented to determine the characteristics of individual spots including size, intensity, blurriness and roundness all of which were used to determine individual spot feature characteristics. Significantly, we determined that spots maintain similar characteristics irrespective of the RNA biomarker and/or tissue used. The verification on 31 microscope images shows agreement of  $R^2 = 0.99$  and a concordance correlation coefficient (CCC) = 0.99 for the total spot counts identified by the observer (115,154) and the algorithm (112,809). We have leveraged the unique detection features of the RNA ISH technology to develop a new method to quantify RNA signal while maintaining tissue context. It is anticipated that this method will enable analysis of gene expression changes in heterogeneous cancer and normal cells and tissues with single cell resolution.

**Keywords:** In situ hybridization (ISH) · Digital pathology · RNA · Image analysis · Quantitative analysis · Image processing · Cancer · Histopathology imaging

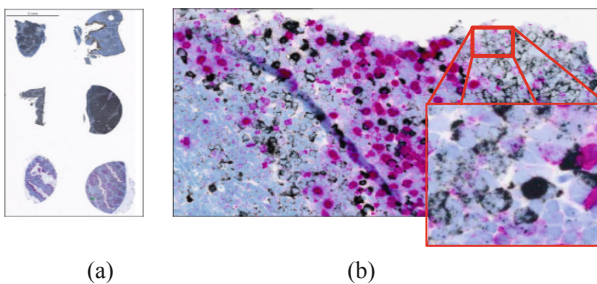
## 1 Introduction

*In-situ* hybridization (ISH) can be used to look for the presence of a genetic abnormality or condition such as amplification of cancer causing genes specifically in cells that, when viewed under a microscope, morphologically appear to be malignant. Unique nucleic acid sequences occupy precise positions in chromosomes, cells and

tissues and ISH allows the presence, absence and/or amplification/expression status of such sequences to be determined without major disruption of the sequences. ISH employs labeled deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) probe molecules that bind to a target gene sequence or transcript to catalyze detection or localization of targeted nucleic acid genes within a cell or tissue sample [1].

Historically, the clinical evaluation of proteins and nucleic acids in tissue has relied upon *in situ* immunoenzymatic detection (staining) methods. For example, detection of B cell clonality is useful for assisting in the diagnosis of B cell lymphomas and such an assessment can be accomplished through the evaluation of KAPPA and LAMBDA light chain expression. As seen in Fig. 1, tonsil tissue stained for KAPPA mRNA may be detected using a black chromogen (silver, Ag) and LAMBDA mRNA may be detected using a purple chromogen (tyramide-sulforhodamine). The presence of the signal of interest appears as tiny spots (e.g. discrete dots) and these spots may accumulate to form larger regions of aggregate signal (hereinafter “signal aggregate blobs” or “blobs”) depending on the expression level (copy number) of each targeted mRNA in B cells. By way of example, plasma cells have approximately 100,000 mRNA copies per cell, and therefore signal in those cells may appear as blobs.

Quantitative ISH analysis will likely be useful in clinical evaluation of a variety of RNA biomarkers; however, its utility remains uncertain due to limitations of existing technologies. An automated technique for estimating an amount of isolated dot signal and signal aggregate blob may facilitate enhanced clinical interpretation of stained biological samples, enable samples to be interpreted more quickly and accurately, and empower evaluation of RNA biomarker clinical utility. In this study, we have developed an image-analysis system and method that enables the detection and quantification of the number of nucleic acid signals present in stained samples.

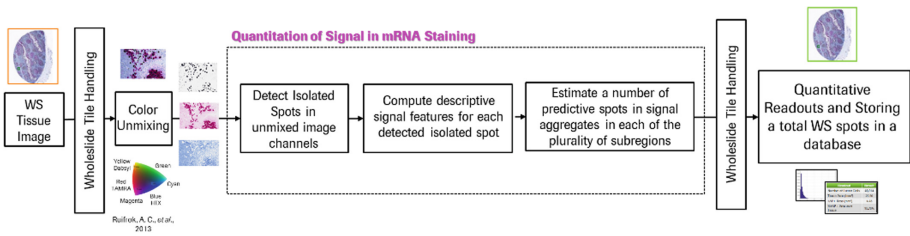


**Fig. 1.** The example of tonsil stained using *in situ* hybridization (ISH) illustrating KAPPA mRNA detected with silver (Ag) in the black color and LAMBDA mRNA detected with tyramide SRB in the purple color: (a) the wholeslide image with six tonsil regions and (b) a field-of-view image at 40X. (Color figure online)

## 2 Methods

The proposed image-analysis framework for detecting and quantifying the expression of the RNA targets (biomarkers) used in our study is shown in Fig. 2.

In this study, we propose a method of estimating an amount of signal corresponding to at least one biomarker in an image of a biological sample comprising: (1) detecting isolated spots in an image (e.g., an unmixed image channel image corresponding to signals from a biomarker); (2) deriving an optical density value of a representative isolated spot (e.g., based on computed signal features or characteristics from the detected isolated spots); (3) and estimating the number of predictive spots in signal aggregates in each of the sub-regions based on the derived optical density value of the representative isolated spot. The method further includes calculation of a total of number of spots in a sub-region by combining a number of detected isolated spots and the estimated number of predictive spots in signal aggregates in each of the sub-regions. Finally, a total number of detected isolated spots combined (i.e. summed) with the estimated number of predictive spots for each sub-region of signal aggregates for the entire tissue slides can be calculated and stored in a database [1].



**Fig. 2.** Image-analysis flowchart illustrating the steps to detect and quantify the expression of RNA targets (biomarkers) in a whole slide image (WS – wholeslide).

### 2.1 Tissue Staining and Digital Images

Using 2.5- $\mu\text{m}$  formaldehyde fixed-paraffin embedded (FFPE) tissue sections, a total of 189 field-of-view (FOV) microscope images and a total of 31 tissue slides of tonsil, lymphoma, and Calu-3 (xenograft) were included in the algorithm development. Tissue slides were stained with a simplex (one color)- and a duplex (two-color)-ISH protocol using probes targeting GAPDH, KAPPA, MALAT1, and KAPPA/LAMBDA RNA transcripts. The staining process was performed using a VENTANA Benchmark Ultra autostainer. All slides were counterstained with Hematoxylin (HTX) in blue color. The 31 slides were scanned using a VENTANA DP 200 scanner. RGB images were obtained with a resolution of  $0.25 \times 0.25 \mu\text{m}^2$  and a typical size of 3 billion pixels or  $20 \times 20 \text{mm}^2$ .

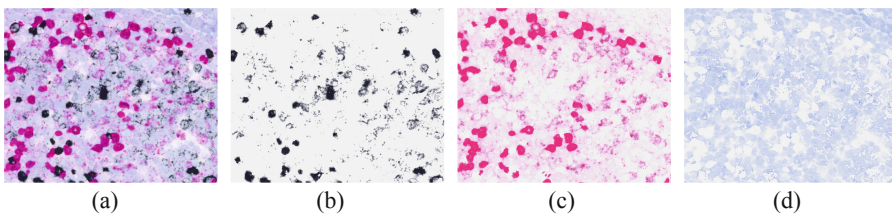
## 2.2 Pre-processing of Color Unmixing

Preprocessing of a color unmixed is performed using a conventional color-deconvolution method to separate different chromogens e.g., black, purple, and blue. In our study, the approach proposed by Ruifrok *et al.* [2] was selected. The unmixing method can be applied to singleplex stained images with one chromogen and counterstain, or applied to multiplex staining images with more than one chromogen and counterstain, as shown in the examples in Fig. 3.

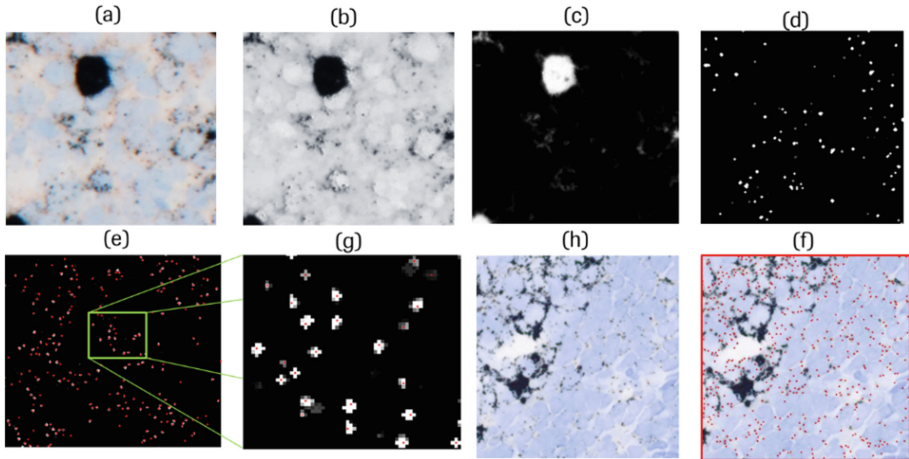
## 2.3 Isolated Spot Detection

Following image acquisition and/or unmixing, an image having a single biomarker channel is provided to the spot detection module such that isolated spots within the image may be detected (as opposed to the “blobs” or aggregate dot signals). An unmixed image channel image is used for input for the spot and blob detection module. A morphological operation is performed to detect isolated spots, i.e. dots, within the image.

As seen in Fig. 4, following the detection of each of the isolated spots in the input image, the detected isolated spots are separated from the blobs in the input image, providing an “isolated spots image channel” and a “blob image channel”. The detected spots are masked out from a blob image channel. In an isolated spots image channel, small objects or blurred point sources can be detected using a multiscale Difference of Gaussians (DoG) approach. Multiple spot sizes are configured in ascending order (small to large), but the processing is in the order of large to small spots. In each iteration, a DoG filter is created from the given inner and outer filter sizes [3]. The respective detections are collected in a resulting seed/annotation object to become the location of each of the detected isolated spots in the  $(x, y)$  coordinates; this location corresponds to the seed center of each detected isolated spot. A seed center can be calculated by determining a centroid or center of mass of each detected isolated spot.



**Fig. 3.** (a) A portion of a whole slide image stained using an *in situ* hybridization assay to detect KAPPA mRNA (black color) and LAMBDA mRNA (purple color) with counterstain hematoxylin; (b, c) an example of an image channel image after unmixing, showing only signal corresponding to KAPPA mRNA (black color) and LAMBDA mRNA (purple color), respectively, and (d) an example of an image channel after unmixing, showing the hematoxylin channel (blue). (Color figure online)



**Fig. 4.** (a) Provides an example of a portion of a whole slide image stained in an ISH assay, (b) illustrates the result of the unmixing of (a) into a single channel (black channel); (c) illustrates a blob channel image whereby the signals from the detected isolated spots from (d) are masked out; (d) illustrates the result of dot detection (a spot channel image) on the unmixed image channel image of (b); (e, f) illustrates derived  $(x, y)$  locations of the detected isolated spots in the spot channel image; and (g) and (h) illustrate an overlay of the detected isolated spots superimposed on the portion of the whole slide image.

#### 2.4 Descriptive Signal Features for Each Detected Isolated Spot

With reference to Fig. 5, the optical density derivation module first computes descriptive signal features for each of the detected isolated spots in the image. The signal feature derivation module implements a Gaussian fitting technique to analyze and parameterize certain characteristics of the detected isolated spots. The fitting method is performed based on the assumption that the distribution of the optical density and the radius is the normal distribution. A 1D-Gaussian-function fitting method is used to estimate the associated spot parameters within a pre-defined patch size surrounding a detected and isolated spot. The patch size is  $7 \times 7$  pixels, which was determined to be the most appropriate patch size for any particular application that will facilitate the provisioning of optimal histogram results.

The characteristics derived from the Gaussian fitting technique include the size, intensity, blurriness, and roundness of the detected isolated dots, and each of these characteristics are computed using parameters of the Gaussian function. By solving the linear system  $\mathbf{Ax} = \mathbf{b}$ , the estimated parameters from the fitting method consist of mean, standard deviation (SD), and full-width-at-half maximum.

By fitting the parameters using the Gaussian model, the computed descriptive signal features of each isolated spot were obtained as following:

1. *Intensity* – is computed using the 98 percentile within the radius of the 5 pixel surrounded the center of the detected spots [no unit].
2. *Blurriness* – refers to the standard deviation ( $\sigma$ ) of the Gaussian-function fitting method.

3. *Size* - refers to the full width at half maximum (FWHM) computed by:

$$FWHM = 2\sqrt{2\ln_2\sigma} \approx 2.355\sigma \quad (1)$$

4. *Roundness* – is the characteristic computed based on the comparison between the actual optical density distribution within a patch and the perfect Gaussian model computed from the estimated parameters. The concordance correlation coefficient (CCC) (which measures the agreement between two variables, e.g., to evaluate reproducibility or for inter-rater reliability) was used to compare the relationship (or the agreement), where  $CCC = 1$  shows that the estimated parameters are perfectly agreement to the ideal Gaussian model; whereas,  $CCC = 0$  shows that there is no agreement between the estimated parameters and the ideal Gaussian model [no unit].

Next, histograms can be generated for each computed signal feature characteristic, as shown in Fig. 5.

## 2.5 Estimation of a Number of Predictive Spots in Signal Aggregates in Each of the Subregions

The generated histograms provide for an understanding of the density of detected isolated cells that have particular values or representative characteristics. The generated histograms therefore provide insight into the characteristics of a representative or typical detected isolated spot. For example, from the intensity histogram (e.g. Fig. 5), it is possible to determine the intensity value of the detected isolated spots that is repeated most often (i.e. the mode of the intensity values). The representative or typical detected isolated spot is then assigned that particular determined intensity value.

The characteristics of the isolated spot representative are used to estimate the number of the spot in the aggregate signals. The estimation assumes a linear relationship between the summation of the optical density for the single spots and the aggregate signals, as following:

$$N = \frac{\sum OD_A}{\sum OD_S}, \quad (2)$$

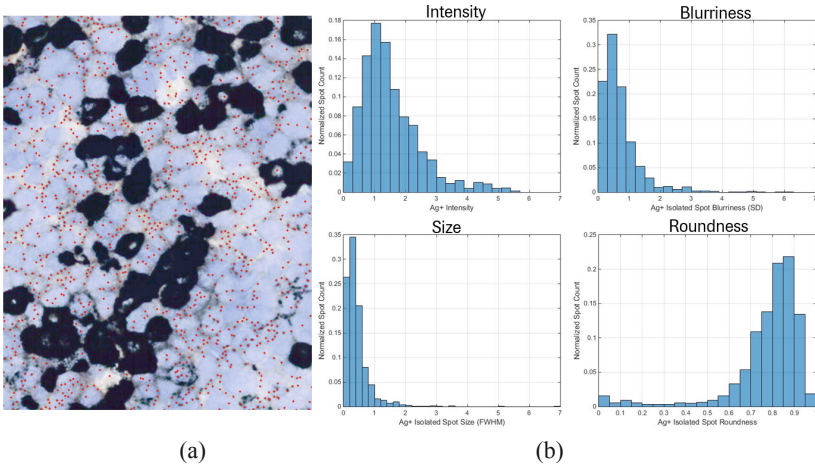
where  $N$  is the number of the spots within an aggregate signal region,  $OD_A$  is the optical density of the aggregate signals, and  $OD_S$  is the optical density of the representative isolated spot signals.

Using the feature histograms of the isolated spots in the previous step, we can apply the individual spot properties in the calculation of their summation of the optical density. The selected properties can be the mode of the intensity (optical density) and the mode of the radius in the feature histograms to calculate the summation of a representative individual spot:

$$\sum OD_S = Area \times \overline{OD_S}, \quad (3)$$



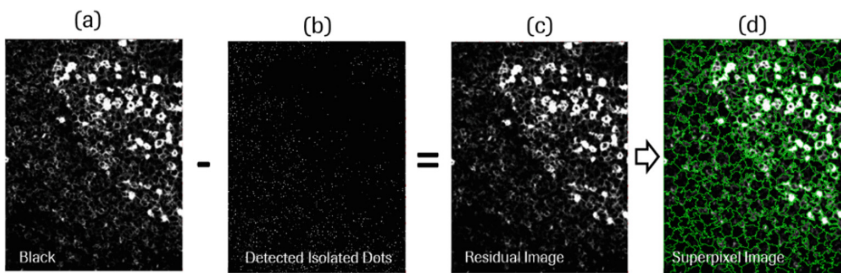
where *Area* refers to a circle ( $\pi r^2$ ) or a rectangle ( $w \times h$ ) area assumed to be a shape of a spot, and  $(\overline{OD_s})$  refers to the representative optical density of a single dot. This can be the mode of the intensity histogram, the average of the total intensity of the total detected isolated spots, or the weighted intensity, etc.



**Fig. 5.** The characteristics of (a) the isolated spots shown in red dots and (b) the feature histograms of intensity, blurriness, size, and roundness, respectively.

**2.5.1 Segmentation and Residual Image Generation**

Prior to estimating the number of predictive spots in signal aggregates, the input image is segmented into a plurality of sub-regions using segmentation. The generation of sub-regions is used to minimize the computation error due to the fact that the computations are based on a smaller local region rather an entire image. The segmentation also reduces the complexity in computing the spot counting in the aggregate signals and the sub-region concept is useful for the quality control verification by an observer and to reduce the complexity in estimating signal in the aggregate signal blobs.



**Fig. 6.** (a) Illustrates the result of the unmixed image in a single black channel, (b) illustrates the detected isolated dot image, (c) illustrates the residual image after masking out the black channel image from detected isolated spot image, and (d) the superpixel segmentation method was applied to the residual image (c).

As shown in Fig. 6, the residual signal is computed by masking out the black-channel image with the isolated spot image. On the residual image, irregularly sized sub-regions can be created by a superpixel segmentation method [4]. The sub-regions of the residual channel image are segmented and grouped the clump signals into smaller regions. Using the superpixel segmentation method, it groups the pixels substantially uniform and perceptually meaningful. The sub-regions using superpixels support in efficient estimation of the number of the signals efficiently. Because some sub-regions have little aggregate signal, it is easy to verify the estimated spot count within that segment. On the other hand, some sub-regions segmented by the superpixel method have completely aggregated signals within the segment, so that it creates a consistent approximation of the spot count within that segment.

Finally, the derived intensity parameter is multiplied by the area to give the optical density of a representative isolated spot. The computed optical density of a representative spot is then supplied to the spot estimation module. Once the number of predictive spots in each sub-region is estimated, the data may be stored in a database or other storage module.

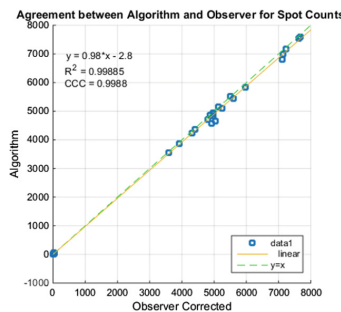
### 3 Results

#### 3.1 Verification of Detected Isolated Spot Counts

The quality control was performed based on a graphic user interface (GUI) which the detected isolated spots overlaid on the original and the observer could correct e.g. add, delete, move the spots. The verification was performed using 31 FOV on the simplex silver microscope images by a trained observer. The agreement plot is shown below with the  $R^2$  of 0.99 and  $CCC = 0.99$ . The example of the spot counting results before and after the correction is in Fig. 7. The correspondence of total spot count identified by the observer (115,154) and the algorithm (112,809) is illustrated in the accompanying Table 1.

**Table 1.** The total spot counts between the algorithm and the observer

Result	Total spot count
Observer	115,154
Algorithm	112,809

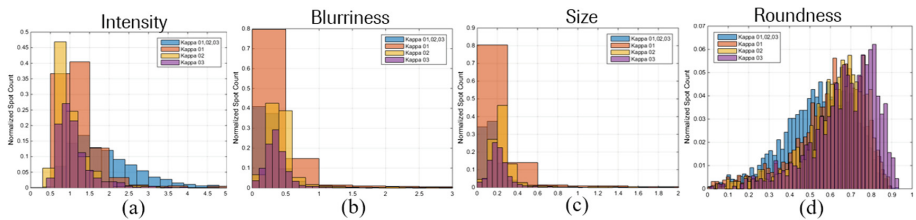


**Fig. 7.** Illustrates the overall scatter of the spot count correspondence between the expert observer and the algorithm results ( $R^2 = 0.99$ ,  $CCC = 0.99$ ) verified on 31 FOVs.

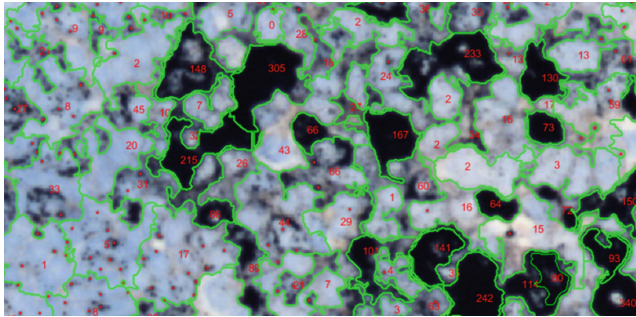


### 3.2 Individual Spot Feature Characteristics and Number of Predictive Spots in Signal Aggregates

We characterized and compared the dots generated by a single probe (i.e., Kappa 01, Kappa 02, or Kappa 03) versus a cocktail of three probes (e.g., Kappa 01, 02, 03), and no probe control using tonsil tissue. As seen in Fig. 8, the intensity of three probes shows wider range than in the one probe images, whereas, the blurriness, size, and roundness characteristics of the spots generated by one probe are not different to spots generated by three probes. As seen in Fig. 9, the analysis result image overlaid with superpixel outlines (green), the overlaid red dots indicating the isolated spots detected by the algorithm, and a red number indicating the additional spots estimated for the aggregate signal within each superpixel.



**Fig. 8.** Illustrates histograms of the spot characteristics of (a) intensity, (b) blurriness, (c) size, and (d) roundness generated by a single probe (i.e., Kappa 01, Kappa 02, or Kappa 03) versus a cocktail of three probes (Kappa 01,02,03).



**Fig. 9.** The analysis result image overlaid with superpixel outlines (green), the overlaid red dots indicating the isolated spots detected by the algorithm, and a red number indicating the additional spots estimated within the aggregate signal with each superpixel. (Color figure online)

## 4 Conclusions

In this study, we have leveraged the unique detection features of the RNA ISH technology to develop a new method to quantify the RNA signal in FFPE tissue, while maintaining tissue context. It is anticipated that this method will enable analysis of gene

expression changes in heterogeneous cancer and normal cells and tissues, with single cell resolution, thereby enabling evaluation of the clinical utility of the plethora of RNA biomarkers encoded in the human genome.

## References

1. Wang, F., et al.: RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Dia.* **14**(1), 22–29 (2012)
2. Ruifrok, A.C., Johnston, D.A.: Quantification of histochemical staining by color deconvolution. *J. Chem. Inf. Model.* **53**(9), 1689–1699 (2013)
3. Polakowski, W.E., et al.: Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency. *IEEE Tran. Med. Img.* **16**(6), 811–819 (1997)
4. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-art superpixel methods. In: *Pattern Analysis and Machine Intelligence* (2012)