

# A Tool for the Quantitative Spatial Analysis of Mammary Gland Epithelium

Rodrigo Fernandez-Gonzalez<sup>1,2</sup>, Carlos Ortiz de Solorzano<sup>1</sup>

<sup>1</sup>Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

<sup>2</sup>Joint Graduate Group in Bioengineering, UC Berkeley/UC San Francisco, CA, USA

**Abstract**—In this paper we present a method for the spatial analysis of complex cellular systems based on a multiscale study of neighborhood relationships. A function to measure those relationships,  $M$ , is introduced. The refined Relative Neighborhood Graph is then presented as a method to establish vicinity relationships within layered cellular structures, and particularized to epithelial cell nuclei in the mammary gland. Finally, the method is illustrated with two examples that show interactions within one population of epithelial cells and between two different populations.

**Keywords**—Mammary gland, quantitative, spatial distribution.

studied. For example, the expression and co-localization of the estrogen (ER) and the progesterone (PR) receptors in luminal epithelial cells [4]; the fact that proliferation markers are not expressed by ER<sup>+</sup> cells [5]; or the possible presence of a niche -a highly organized pattern of different cell types- around mammary stem cells [6] have never been assessed from a quantitative, spatial point of view.

In order to address these questions, we have developed a quantitative spatial analysis tool that we have integrated into an existing 3D microscopy system [7]. In this paper we introduce that tool and show two examples obtained on real data, thus demonstrating how we will use it to address some of the questions mentioned above.

## I. INTRODUCTION

As it is already accepted in our current systemic approach to biology, the development, function and regeneration of tissues, both normal and abnormal, is determined by the location and distribution of various cell types and phenotypes within the tissue [1, 2]. For example, in mammary gland development, cap cells with invasive properties line the surface of the growing ducts, which extend through the fat pad that embeds the gland. In the alveolar units that cap the ducts of a mature gland, secretory epithelial cells line the lumen of the ducts. After pregnancy, this mature luminal epithelium secretes milk into the ducts. Myoepithelial cells are arranged around the ductal tree, as well as around the terminal alveolar units. Upon, the appropriate stimulus they contract, thus forcing the milk through the ducts towards the nipple. Even within a given cell type or subtype, heterogeneous cellular phenotypes are required to maintain the proper tissue homeostasis. Thus, hormone receptor sensitive ductal epithelial cells, supposedly responsible of cellular signaling, are interspersed with hormone receptor negative cells, which carry out the functions dictated by the hormone sensitive ones [3].

Many of these spatial phenomena have been previously described in a qualitative way but, due to the lack of appropriate tools, most of them have not been quantitatively

## II. METHODOLOGY

### A. Tissue processing

Paraffin-embedded mammary gland tissue blocks are sectioned at 5  $\mu$ m, and the sections are immunostained for the appropriate antigens. A counterstain is used that allows us to study the morphology of the tissue (e.g.: DAPI). Low magnification (2.5X) images of the counterstaining of all the sections are then automatically acquired using a motorized Zeiss Axioplan I microscope coupled with a monochrome XilliX Microimager CCD camera. This is done automatically by scanning the area of the slide occupied by the tissue; correcting the microscope focusing whenever necessary; and tiling together all the individual snapshots into a single image of the entire section.

The next step consists of automatically annotating the structures of interest (ducts, tumors, ...) in these low magnification images [8]. The annotated structures are then used to create a three-dimensional model of the sample. With this model we can track the morphology of the tissue to determine which are the areas where a spatial analysis might be more interesting. After selecting these areas, the system asks the user to place the right fluorescent section(-s) on the stage, and high magnification (40X) images of the immunostaining of the chosen areas are acquired. In these images nuclei are manually annotated with dots and visually classified, forming a point pattern; they can also be automatically segmented and quantified [9]. In the later case, the center of mass of each nucleus is computed to obtain a point pattern of nuclei markings.

### B. $M$ function analysis

#### B.1. Definitions

Given a set of points  $\{n_1, \dots, n_{N_c}\}$  representing the nuclei belonging to population  $C$  in the study area, we define  $n_{iC_r}$ ,

---

This work was supported by the U.S. Army Medical Research Materiel Command (grants DAMD17-00-1-0306 and DAMD17-00-1-0227), the Lawrence Berkeley National Laboratory Directed Research and Development program, and the California Breast Cancer Research Program (project 8WB-0150). R Fernandez-Gonzalez was supported by a predoctoral fellowship from the Department of Defense Breast Cancer Research Program (BC020294).

the number of neighbors of nucleus  $n_i$  belonging to population  $C$  within distance  $r$ ;  $n_{ir}$ , the total number of neighbors of  $n_i$  (belonging to any population) within distance  $r$ ;  $N_C$ , the total number of nuclei belonging to population  $C$  within the area under study; and  $N$ , the total number of nuclei in that same area.

### B.2. Single-variable analysis

Mammary gland epithelial cells are located at ducts, end buds and alveolar structures, but never in the surrounding stromal and fatty tissue. For that reason, we cannot do our spatial analysis with Ripley's  $K$  function [10],-traditionally used for this task in other fields-, since it assumes that the cells can be located anywhere in the image space. Thus, we now assume a space with constrained nuclear locations. In this space, the total number of nuclei on an area measures the size of the set of possible locations for an epithelial nucleus in that area: any of those points could be occupied by a nucleus. Thus, we can measure the density of nuclei belonging to population  $C$  as a ratio of nuclei numbers, and so, we define [11]:

$$M(r, C) = \frac{\sum_{i=1}^{N_C} \frac{n_{iCr}}{n_{ir}}}{N_C} / \frac{N_C}{N} \quad (1)$$

The numerator of (1) computes the average density of neighbors belonging to population  $C$  within distance  $r$ , and then compares that value to a benchmark: the density of nuclei belonging to population  $C$  in the entire study area. Therefore, clustered patterns of nuclei will have  $M(r, C) > 1$ , with a peak at the cluster size. On the other hand, regular patterns will have  $M(r, C) < 1$ . Finally, random distributions will have  $M(r, C) = 1$ . In general, we can say that  $M(r, C) = k$  implies that the density of nuclei belonging to population  $C$  within distance  $r$  is  $k$  times that of the entire area under study.

To complete the univariate analysis we need to have a way to establish the significance of our measurements. For that reason, we run  $m$  Monte Carlo simulations of the nuclei distribution within the area of interest. The simulations are set up by preserving the nuclei locations and randomly assigning the population each nucleus belongs to. We compute the  $M$  function for each one of these simulations ( $M_i(r, C)$ ,  $i = 1, \dots, m$ ) and calculate  $U(r, C)$  and  $L(r, C)$ :

$$U(r, C) = \max_{i=1, \dots, m} (M_i(r, C)) \quad (2)$$

$$L(r, C) = \min_{i=1, \dots, m} (M_i(r, C)) \quad (3)$$

Now we can plot  $M(r, C)$ ,  $U(r, C)$  and  $L(r, C)$  in the same graph. Peaks of  $M(r, C)$  above  $U(r, C)$  are evidence of significant clustering (with confidence level  $\alpha = 1/(m + 1)$ ). Similarly, troughs below  $L(r, C)$  represent significant regularity or dispersion. Any nuclei distribution with no significant peaks or troughs can be considered to be random.

### B.3. Multiple-variable analysis

The  $M$  function analysis described in the previous section can be used to study the distribution of cells within a given population, e.g. Expressing or not a given marker. However, most of the problems introduced in section I involve two or more cell populations. In order to study this type of problems, we can modify (1) to get:

$$M(r, C_1, C_2) = \frac{\sum_{i=1}^{N_{C_1}} \frac{n_{iC_2r}}{n_{ir}}}{N_{C_1}} / \frac{N_{C_2}}{N} \quad (4)$$

where  $C_1$  and  $C_2$  are the two populations under study,  $N_{C_1}$  and  $N_{C_2}$  are the number of nuclei in each one of those populations and  $n_{iC_2r}$  is the number of nuclei belonging to population  $C_2$  within distance  $r$  of nucleus  $n_i$  (with  $n_i$  belonging to  $C_1$ ). It is easy to see how these equation could be extended to three or more populations.

Now  $M$  values larger than 1 are indicative of attraction between populations  $C_1$  and  $C_2$  (a special case of this is *co-localization*, i.e., attraction at distance  $r = 0$ ); values smaller than 1 indicate repulsion; and  $M(r, C_1, C_2) = 1$  shows independence of the spatial distributions of both populations at distance  $r$ . However, significant values of  $M(r, C_1, C_2)$  maybe due either to actual interactions between both populations or to the patterns of each one of them. For this reason, we set up our Monte Carlo simulations preserving the locations of the nuclei belonging to population  $C_1$  and redistributing the location of population  $C_2$ . Thus, we control for the  $C_1$  pattern. The same process is applied to  $M(r, C_2, C_1)$ . Finally, significant interaction at distance  $r$  is only accepted if both  $M(r, C_1, C_2)$  and  $M(r, C_2, C_1)$  are significantly different from randomness.

### C. Refined RNG

In the previous section we have used the shortest Euclidean distance to measure how far apart two nuclei are. However, this is not the best way to represent vicinity. In fact, the shortest Euclidean distance is some times obtained by traversing luminal areas, thus defining neighborhood relationships that may have topological meaning, but that cannot explain the underlying cellular interactions that we wish to measure with our analysis, since cell-to-cell signaling in the epithelium normally occurs through intermediate cells [12]. Therefore, we decided to model our tissue using a graph where the nodes are the different nuclei, edges represent neighborhood relationships and distances can be measured as the number of edges between two nuclei.

We start out by building a Delaunay triangulation using the nuclei markings as nodes. This provides a preliminary tessellation where we can already measure distances as number of edges. On top of this triangulation we can now build a Relative Neighborhood Graph (RNG). Here, we

preserve an edge if and only if the two nuclei on its sides ( $n_i$  and  $n_j$ ) are relatively close [13], that is:

$$d(n_i, n_j) \leq \max(d(n_i, n_k), d(n_j, n_k)) \quad (5)$$

$$\forall k=1, \dots, N, k \neq i, j$$

where  $d(n_i, n_j)$  is the length of the edge between  $n_i$  and  $n_j$ . In other words, what this definition states is that an edge in the Delaunay triangulation is preserved if the nuclei on its sides are at least as close to each other as they are to any other nucleus in the graph. With this we obtain the RNG. Finally, we do a refinement step where we get rid of all the edges which are too large and we force connections between nuclei which are too close (using the shortest Euclidean distance) to not to be neighbors. Now, using Floyd's or Dijkstra's algorithms [14], we can easily build a table with the shortest distance (measured as the number of edges) between each pair of nuclei in the graph.

### III. RESULTS

We ran our spatial analysis tool on a set of synthetic images to test for its accuracy at detecting interactions. Then we went on to do the analysis of real tissue samples. In this section we describe two different examples. For the first one the tissue was obtained from a transgenic mouse overexpressing the HER2 gene, a growth factor receptor whose human counterpart is overexpressed in about 30% of breast cancers. Sections were taken from this sample and immunostained for HER2 using diaminobenzidine (brown precipitate). The nuclei were counterstained with hematoxylin (blue). High magnification images of certain areas in these sections were acquired, and the nuclei in those areas were manually annotated. Fig. 1 shows the spatial analysis of the HER2<sup>+</sup> population in one of those areas (inset). Positive nuclei are marked with red dots, negative ones are green. The analysis indicates the presence of clustering at small distances around the nuclei (at 2 to 8 nuclei of distance as measured by edges in the refined RNG), with a peak at distance  $r = 2$ . This peak reveals the presence of clusters of HER2<sup>+</sup> cells with a radius of 2 nuclei and a density of  $M = 1.55$  ( $\alpha = 0.05$ ).

For the second example we obtained tissue from a wild type mouse which had been given a constant dose of BromodeoxyUridine (BrdU) for two weeks. BrdU is a thymidine analog that gets incorporated into the DNA of the cells that undergo mitosis. The tissue was sectioned, and double immunofluorescence staining was carried out on the sections. BrdU was detected using a secondary antibody labeled with Alexa 568, a red fluorochrome, while Alexa 488 (green) was used to detect ER<sup>+</sup> cells. Nuclei were counterstained with DAPI (blue). Once again, high magnification images of some areas were taken, and nuclei were manually annotated. Then, multiple-variable analysis of the interaction between ER<sup>+</sup> and BrdU<sup>+</sup> cells was carried out. Fig. 2 shows one of the analyzed areas, where nuclei

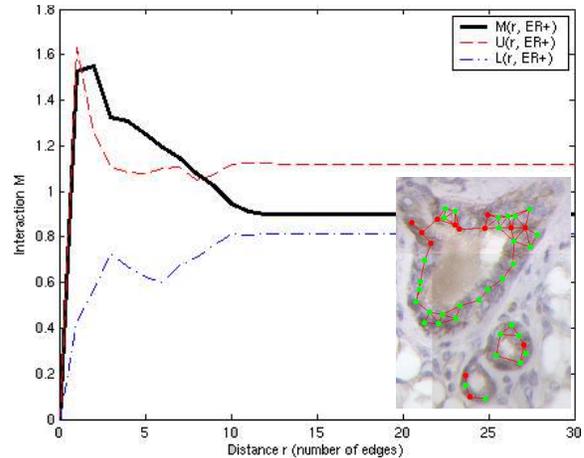


Fig. 1: Single variable spatial analysis of HER2<sup>+</sup> nuclei.

markings have been removed for clarity. The refined RNG establishing neighborhood relationships is also shown. Fig. 3 contains the results of the analysis ( $M(r, ER^+, BrdU^+)$ ) for this area. The graph for  $M(r, BrdU^+, ER^+)$  is very similar to the one shown here. Thus, there seems to be repulsion between both populations at very small scales. Actually, the peak of this repulsive interaction is at distance  $r = 0$  ( $M = 0.15$ ,  $\alpha = 0.05$ ), i.e., ER<sup>+</sup> and BrdU<sup>+</sup> cells do not co-localize most of the times, but do co-localize occasionally. This result, whose intensity and extent we can now quantify using the  $M$  values, has previously been described qualitatively in the literature.

### IV. DISCUSSION

Studying the function of complex biological systems requires the quantitative analysis of heterogeneous cellular populations located in diverse topological patterns. In order to do this in a way that provides consistency and high throughput, quantitative tools for the spatial analysis of samples are required. In this paper we have presented a method that automatically provides a measurement of the way cells interact within one population, as well as of the different types of interaction that might occur between the different cell populations present in a tissue. Our approach is based on a multiscale analysis of the number of neighbors belonging to the population under study, followed by comparison to a benchmark, the total density of nuclei within that population in the entire study area. Thus, we define the  $M$  function, which takes into account the

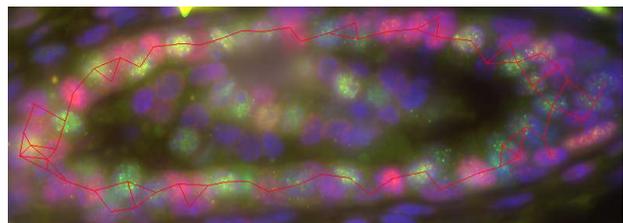


Fig. 2: ER<sup>+</sup> (green) and BrdU<sup>+</sup> (red) cells in a duct.

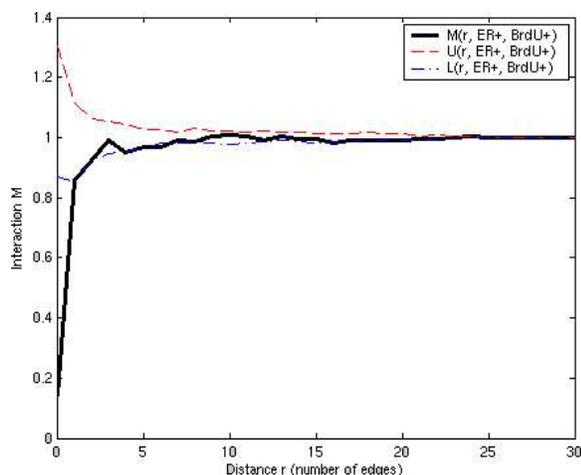


Fig. 3: Spatial analysis of the interaction between  $ER^+$  and  $BrdU^+$  nuclei.

heterogeneous distribution of the epithelium within mammary tissue. This function, -together with the analysis scheme where it is embedded-, allows for unsupervised analysis of large data sets in a reasonable time, since it does not include any complex calculation. The method provides comparability of concentration measurements across populations; remains unbiased concerning different scales; and can be modified depending on the desired significance level.

In order to define neighborhood in a way that takes into account the histology of the tissue as well as differences in cell size/image magnification, we create a refined RNG that has the nuclei markings as its nodes. The connections in this graph represent vicinity in a way that faithfully depicts what nuclei might be directly interacting with each other in the tissue.

In the near future we are planning on using this tool to address several problems, including co-localization/interaction studies of  $ER^+$ ,  $PR^+$  and  $HER2^+$  populations in both wild type and transgenic mice, or characterization of the distribution of label-retaining cells (a population of cells likely to be enriched for mammary stem cells) with respect to other populations present in the mammary epithelium, thus trying to unveil the presence of a niche around stem cells similar to the ones observed in other organs. Since both of these problems are inherently three-dimensional, we are currently working on adding one more dimension to our analysis scheme. This extended functionality should help us explore in further detail the possible determinants of interaction both within and between cell populations.

## V. CONCLUSION

In this paper we have presented a method for the spatial analysis of mammary epithelium. Thus, we have created a tool to quantitatively measure what previously could only be

qualitatively described. Our multiscale method is consistent, comparable across populations and allows for automatic, high-throughput analysis of large data sets. The use of this approach to study problems where interactions between cells are expected will greatly contribute to the detailed description of these phenomena.

## REFERENCES

- [1] E. Fuchs, "Beauty is skin deep: the fascinating biology of the epidermis and its appendages", *Harvey Lect.* 94, pp. 47-77, 1998.
- [2] W. Imagawa, J. Yang, R. Guzman, S. Nandi, "Control of mammary gland development" in *The Physiology of Reproduction 2<sup>nd</sup> Ed.*, E. Knobil, J.D. Neill Eds. New York Raven Press., 1994, ch. 3, pp. 1033-1063.
- [3] M. Smalley and A. Ashworth, "Stem cells and breast cancer: a field in transit", *Nat Rev. Cancer*, vol 3, no.11, pp. 832-844. Nov. 2003.
- [4] G. Shyamla, Y.C. Chou, S.G. Louie, R.C. Guzman, G.H. Smith, S. Nandi. " Cellular expression of estrogen and progesterone receptors in mammary glands: regulation by hormones, development and aging", *J. Steroid Biochem. Mol. Biol.*, vol. 80, no. 2, pp. 137-148, Feb. 2002.
- [5] R.B. Clarke, A. Howell, C.S. Potten, E. Anderson, " P27(KIP1) expression indicates that steroid receptor-positive cells are a non-proliferating, differentiated subpopulation of the normal human breast epithelium", *Eur. J. Cancer*, vol. 36, suppl. 4, pp. 28-29, Sep. 2000.
- [6] N.J. Kenney, G.H. Smith, E. Lawrence, J.C. Barrett, D.S. Salomon, " Identification of Stem Cell Units in the Terminal End Bud and Duct of the Mouse Mammary Gland", *J. Biomed. Biotechnol.*, vol. 1, no. 3, pp. 133-143, 2001.
- [7] R. Fernandez-Gonzalez, A. Jones, E. Garcia-Rodriguez, P.Y. Chen, A. Idica, S.J. Lockett, M.H. Barcellos-Hoff, C. Ortiz-De-Solorzano, "A system for combined three-dimensional morphological and molecular analysis of thick tissue specimens", *Microsc. Res. Tech.*, vol. 59, no. 6, pp. 522-530, Dec. 2002.
- [8] R. Fernandez-Gonzalez, T. Deschamps, A. Idica, R. Malladi, C. Ortiz-De-Solorzano, " Automatic segmentation of histological structures in mammary gland tissue sections", *J.Biomed. Optics*, in press.
- [9] N. Malpica, C. Ortiz-De-Solorzano, J.J. Vaquero, A. Santos, I. Vallcorba, J.M. Garcia-Sagredo, F. del Pozo, "Applying watershed algorithms to the segmentation of clustered nuclei : defining strategies for nuclei and background marking", *Cytometry*, no. 28, pp. 289-297, 1997.
- [10] T.C. Bailey and A.C. Gatrell, *Interactive Spatial Data Analysis*. Essex, England: Prentice Hall, 1995, pp. 90-95, 103-105, 117-131.
- [11] E. Marcon and F. Puech, "Measures of the geographic concentration of industries: improving distance-based methods", *Cahiers de la MSE*, 18, p. 22, 2003.
- [12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walters, *Molecular Biology of the Cell*, 4<sup>th</sup> ed.. New York, NY: Garland Science, 2002. Ch. 15.
- [13] G.T. Toussaint, "The relative neighborhood graph of a finite planar set", *Pattern Recognition*, vol. 12, pp. 261-268, 1980.
- [14] B.R. Preiss, *Data structures and algorithms with object oriented design patterns in Java*, John Wiley & Sons, 1999. Available: <http://www.brpreiss.com/books/opus5/>