Assessing the Cancer Stem Cell distribution in tumorspheres.

Jerónimo Fotinós[1], **Lucas Barberis**[1,*,+], **and Luciano Vellón**[2,+]

[1]IFEG, FAMAF, CONICET, UNC, Córdoba, Argentina
[2]Stem Cells Lab, IBYME, CONICET, Buenos Aires, Argentina
[*]lbarberis@unc.edu.ar
[+]these authors contributed equally to this work

ABSTRACT

In previous theoretical research, we inferred that Cancer Stem Cells, the cells that presumably drive tumor growth, are not uniformly distributed in the bulk of a tumorsphere. This knowledge is useful for mathematical modeling, which most of the time assumes uniformity as a simplification. Besides, knowing the Cancer Stem Cells' distribution is crucial for the understanding and designing further treatments of cancer disease.
In this article, we present our first experimental results joined with a specific method to process confocal images that allow us to measure the amount and location of Cancer Stem Cells in tumorspheres. We report the complete set of computational and statistical steps to achieve this and compare the experimental results with previous predictions obtained by simulations.

1 Introduction

Cancer Stem Cells (CSCs) are defined by their capacity to self-renew and differentiate to give rise to phenotypically diverse cells. They are resistant to conventional anti-cancer treatment being thus implicated in disease recurrence and metastasis[1,2]. One of the main barriers to the development of CSCs-targeted therapies is their scarcity in vivo, limiting their availability as experimental systems for pharmaceutical development, and raising the need for production at a scale large enough to fulfill academic and industry requirements. A common solution is the anchorage-independent growth of cancer cells to generate a 3D culture of mammary epithelial cells[3], which has been shown to enrich cells with CSCs-like properties since most epithelial cells are killed by anoikis[4]. In this way, taking into account the limitations of this assay *no see a que limitation see refiere*, the resulting tumorspheres are formed by the clonal expansion of a single cell, instead of the self-aggregation of existing cells[5].

Still, routinely used cell culture techniques are material- and labor-consuming tasks that generate a great amount of inter-culture variability and contamination risks. Moreover, traditional cell culture at a large scale is also cost-ineffective in terms of the high investment in cell culture media and growth factors. Thus, developing forefront, high-throughput screening platforms to identify cytotoxic inhibitors and/or differentiation-promoting agents targeting CSCs would require the optimization of a series of bioprocesses, not only to enable the massive culture of undifferentiated cancer cells but also the analysis of high volumes of data.esta ultima oraci'on es largiiiisima

One critical downstream part of the process is to assess the cellular response in terms of viability and/or stemness markers, which requires external software for image analysis and segmentation. Even though high-throughput cytometric methods have been developed[6], there is still the need to know the exact identification, location, and targeting of putative CSCs. This would allow, in turn, modeling CSCs dynamics in the context of well-defined conditions such as tumorspheres and develop cost-effective and predictive tools for the examination of tumor evolution, response to therapy, and therapeutic evaluation of novel anti-CSCs compounds. esta ultima oraci'on es largiiiisima

Mathematics has offered positive prospects in cancer research, being widely used. Many biological problems demand methods and techniques requiring, not only traditionally applied mathematics, but also pure mathematics, statistics, and computation. l aoraci;on queue sigue deberia decir queue funciona en cancer, poner mas referenciasIt success for the modeling of tumor responses (Vieira et al., 2023[7]). Indeed, analytical mathematical methods allowed us to determine the expected fraction of CSCs for tumorspheres in different culture conditions[8–10] and, the effect of specific therapies on their development[11].

We have also computationally simulated the growth of a colony of cells in two dimensions using an Agent-Based Model (ABM) that mimics basic features of CSCs proliferation to form a spheroid[12]. The simulated spheroid grows from a single CSC and the cells can undergo mitosis at a fixed rate (the PDT). Depending on the genetic and environment of the CSC, it will replicate giving birth to another CSC with a certain probability p_s, otherwise it will differentiate giving birth to a *differentiated cell* (DC). This simple model, allowed us to estimate the total number of CSCs, the fraction of CSCs situated on the periphery of the colony, and the size of the whole spheroid, showing that these traits are dependent on the self-renewal probability (p_s) of the CSC population. Indeed, simulating with intermediate self-renewal probabilities for the CSCs, we observed active CSCs at the border of the colony and detected that they form a path, that links the center of the colony with its border. Furthermore, increasing the self-renewal probability of the CSCs led, as expected, to a large CSCs population that overtook the system. This last situation describes most experimental conditions used for culturing tumorspheres[13,14] and agrees with previous mathematical models[8–10].

Inspired by these simulations, we performed tumorspheres assays grown from MCF-7 cells and then took and analyzed their corresponding confocal images. This analysis consists of an advanced segmentation method able to detect the expression and distribution of the stem-like cells (Sox2-positive cells), under the assumption that this stemness factor is expressed in tumorspheres from cell lines and primary cultures, in this particular case, from breast cancer patients (Leis et al., 2012[15]). With these processed images we studied the distribution patterns of CSCs using statistical tools that validate our main computational finding: that CSCs are heterogeneous distributed in a tumorsphere.

qu'e encontramos con enfasis en el metodo y no en el resultado. Decir que el metdo permite unir ambos modelos el computacional y el biológico

Regardless of the non-extensiveness of the analyzed cases, we highlight the method presented here, and its capacity to extract data from the experiments in a way that allows for comparing with the simulated cultures.

Results and discussion

We are interested in the CSCs distribution inside tumorspheres which, according to simulations, must form "paths" connecting each other. For this reason, we performed a tumorsphere assay where MCF-7 cells proliferate in a solution rich in stem growth factors to ensure a large fraction of CSCs in the culture. After 9 days of growth, we collected the spheroids and attached them to a slide by cytospin, thus their spheroidal shape may be lost becoming discs. Then, we stained the slides with the corresponding primary and secondary antibodies needed to mark the location of all cellular nuclei (DAPI) and the position of the stem cells (SOX2). Finally, we took pictures, by means of a confocal microscope, slicing the spheroids and obtaining a full 3D reconstruction of them. These images were processed with a Data Science approach, using techniques of computer vision and statistical analysis, among others. The result of the process allow us to mathematically reconstruct the discs, specifying the position of all their cells, and marking those that are candidates to be CSCs. The experimental and image processing procedures are both fully detailed in the 1 Methods section.

The Cancer Stem Cells distribution.

Our main results are summarized in Fig. 1, in which we can observe the distribution of the two cell phenotypes: CSCs in red and DCs in blue. We present three representative examples belonging to the reconstruction of three different spheroids. After the filtering and reconstruction process of the confocal images, we obtained pictures of the cell's distribution as a Voronoi tessellation. That is, each cell is represented, in Fig. 1, as a polygon whose centroid coincides with the centroid of the cell in the pictures. Note that this representation is just an approximation of the shape of the cell that allows us to quantify the SOX2 content inside the cell and decide, using statistical tools, if the cell belongs to the Stem niche or to the Differentiated kind. After the whole process, the CSCs become represented by red polygons and the DCs by blue polygons. In the following, we will refer to cells indistinctly as CSC or RED and DC or BLUE

Spots in the border

Our first example, depicted in Fig. 1a, is a spheroid that has not grown much. Because the culture medium is not homogeneously distributed, we expect a large inter-spheroid variability in both the number of cells, and the CSCs fraction. In this case, the spheroid had just a hundred cells, with half of them being CSCs. Because the experimental setup is rich in stemness factors, the resulting spheroids are rich in CSCs. However, for spheroids that could have been spread in a single layer, as those depicted in Fig. 1, it is evident that there is a higher concentration of the CSC in the core of the spheroids. This occurs because the first CSC has a high chance of giving birth to another CSC. But, when one of these CSC differentiates the first time, its lineage will only be made by DC cells , depicted in blue. These examples show exactly what we expected according to our simulations: if the first cell is a RED CSC, that drives the spheroid growth, it must be close to the center. With a large non-differentiating probability p_s, we expect a RED core surrounded by BLUE cells, because, when the first CSC differentiated, becoming a BLUE cell, all the cells that will pop up near to it must be BLUE. The only exception is when the BLUE cell cannot undergo mitosis in a close place already occupied by a RED one. Indeed, this RED cell will continue giving birth to RED cells until it is differentiated, leaving in the process a RED path among BLUE cells. This is exactly what is shown in panel (a): at an early stage, RED cells become surrounded by BLUE cells in the lower portion of the spheroid. However, three RED cells were able to initiate paths that extend to the periphery. The result is that CSC will form spots in the border of the spheroid.

Jero

Established mechanism or what we want to prove?

Jero

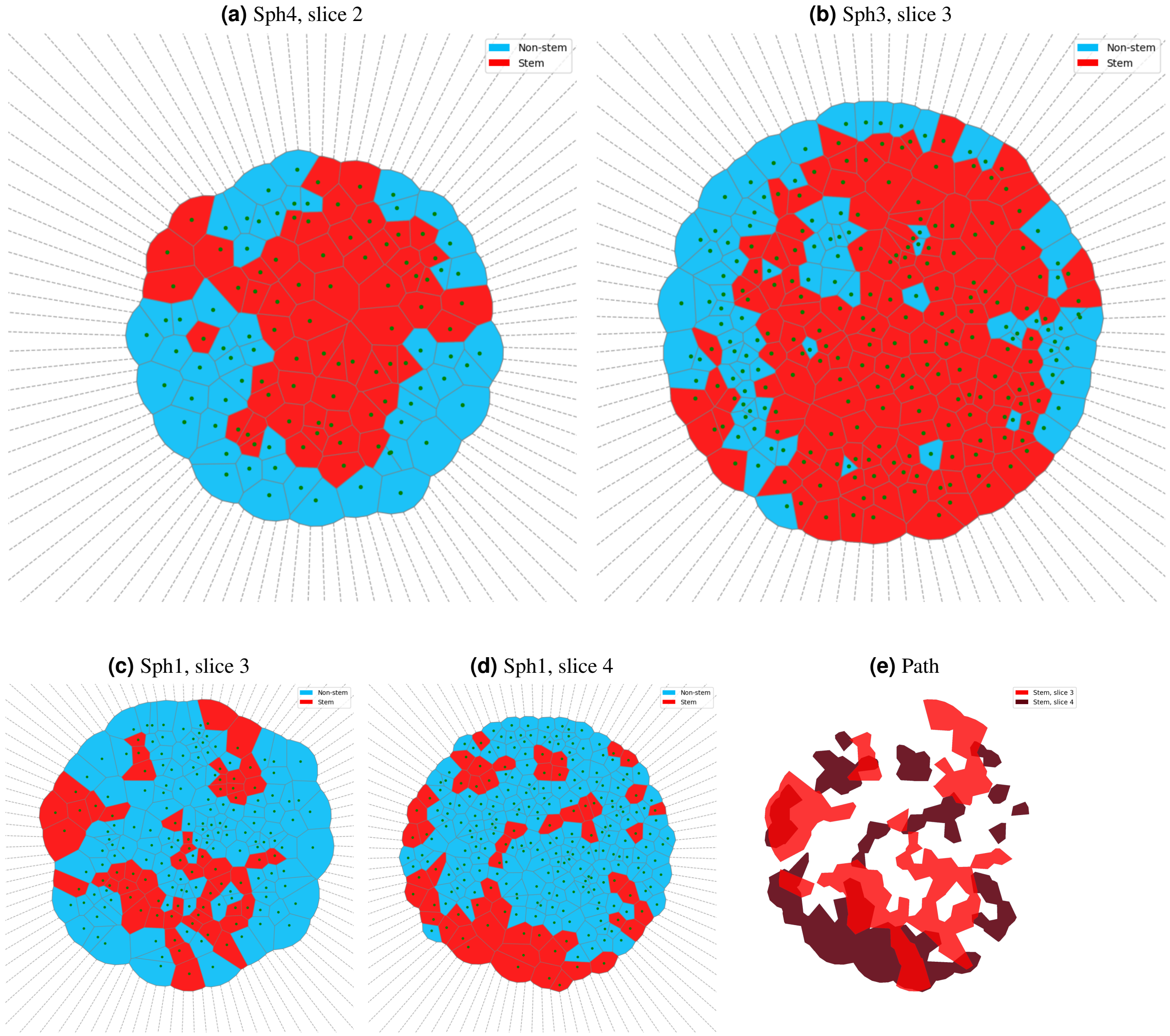
DC cells is kind of redundant, I'd replace everything by

A word of caution is now convenient. We know that the original object is three-dimensional, while here we are observing a 2D projection of it. Thus, it is naturally arguable that the result could be just a coincidence. But, in the process of pressing the spheroid against the slide, it is easy to imagine (and prove) that the symmetry of the spheroid will be broken in the direction perpendicular to the slide (the z-axis). As a consequence, the radial distribution of the cells around the z-axis (that passes through the center of the disc) becomes almost unaltered. Furthermore, this explains why the number of cells in the center of the discs is larger than in their border.

For the case of a larger growth rate, as shown in Fig. 1b we can assume that the fraction of RED cells is larger than in the 54% in example of panel 1a but, the cell new cell fraction is 64% . This result, almost trivial, supports one of the strongest modeling hypotheses of the simulation: that the measured growing rate can only be measured for the bulk of the cells[10] and the outcome of the mitosis of a CSC is a probability that scales such rate[12]. Of course, all the quantitative results reported in Barberis, 2021[12] must be significantly different when we shift from two to three dimensions but, the qualitative ones remain the same: the CSC will form paths and are heterogeneously distributed in the border of spheroids after approximately one week.

Jero

I'm not following.


(a) Sph4, slice 2

(b) Sph3, slice 3

(c) Sph1, slice 3

(d) Sph1, slice 4

(e) Path

Figure 1. cambiar las leyendas: non-stem por differentiated Reconstruction of the disc after image processing. The cells are colored in red for CSC and blue for DCs. It is evident that the CSC (red regions) are not uniformly distributed in any picture, and they seem not to prefer the border of the aggregate **esto see tiene que notar**.

Two layers sample

As mentioned, our method of placing the sample on the slide deforms the spheroids projecting the cells onto a disc. Depending on the forces acting on the spheroids, sometimes their cells can not push all their neighbors apart, piling up on them instead of becoming surrounded . We observed that, for this particular experiment, the resulting discs have two cell layers superimposed. We thus must process two images, placing the focal plane on each of these layers, to reconstruct the spheroid. In panels (c) of Fig. 1 we depict the obtained cell distribution for the upper (c1) and bottom (c2) layers. At first glance, we can not identify the RED core described in the previous subsection. This is another consequence of the heterogeneity of the CSCs distribution. For example, in panel 1a the RED spots were localized in the upper half of the spheroid, and along the z-axis. The RED cells along this axis will be pushed all together, increasing the RED core density as depicted. On the other hand, if the spheroid had been rotated with 90deg respect to a horizontal axis, most border BLUE cells would have gotten among RED ones. Indeed, something like this occurs in the example in panel 1c, where there are BLUE spots in the bulk of the spheroid. In the two-layer disc, we have an example of such a situation where it becomes difficult to follow the CSCs' path. The RED cells look almost randomly distributed but, we have here the opportunity for a better reconstruction of the spheroid because we lost less spatial information. Indeed, if we subtract BLUE cells from snapshots in panels (c1) and (c2) and overlay the resulting RED cell populations, we obtain the cell distribution in panel (c3). Note that we colored the bottom layer cells (against the plate) in dark-red, and in light red the cells in the upper layer. The result is now evident, most of the RED cells are connected, forming a path along the two layers as we expected.

Jero
Somewhat unclear.

Jero
Names of panels do not match.

Non-randomness of the CSC's distribution

With the image processing method reported here, we can obtain information that is highly relevant for setting up simulations, formulating mathematical models, and fitting their results. Indeed, we can measure the total amount of cells in spheroids and the CSCs fraction, this information increases the accuracy of our models[10]. In the upper half of Table 1 we summarize the cell counts for each example. Besides, we now know the growth rate is $r = 1.1\ cell/day$, which is a bit larger than the one we first estimated in our simulations, but indistinguishable from those recovered in our mathematical models[10].

Jero
The difference between sph1 and the rest should be ps, not the projection.

Jero
They don't.

However, even more interestingly, the method gives us the information necessary to quantitatively assess if CSCs are forming groups in a way that is not explained by randomly locating them inside the spheroid. As explained in the methods section (1), for doing this characterization, we build the network that has cells as nodes, and links between the nodes of neighboring cells (c.f. Fig. 6). Then, we test for network homophily: the tendency of cells to neighbor cells of the same stemness. This was characterized by the homophily ratio and the assortativity coefficient (both of which increase with network homophily). These values can be seen in the bottom half of Table 1. To enable meaningful comparison (and hence interpretation) of these values, we turn to a classical approach from the physics of complex systems. We calculated these parameters for a set (ensamble) of graphs where CSCs were distributed randomly in an otherwise identical network. Using an ensamble of size 10000, we were able to perform statistical tests (all with significance $\alpha_t = 0.001$) to see whether the values for the experimental network differed significantly from the ones for the networks with randomly located CSCs. The result of these tests was that both parameters were significantly higher than random for all spheroids. This enables us to conclude that there is a significant tendency for cells in the spheroids to neighbor other cells of the same stemness.

Jero
Do we?

Another way of reaching similar conclusions, is looking at the stem subgraphs. The stem subgraph of a given graph is the network formed by stem cells and connections between them. By looking at the degree distribution of the stem subgraphs for the different spheroids, we can calculate the mean number of CSCs neighboring a given CSC. We can also calculate the number of connected components of this graph, which is the number of separated groups of CSCs. The more clustered the CSCs, the less the connected components. These values can again be seen in Table 1. Once more, we can generate many copies of the network of each spheroid, locating the CSCs in a random way. We can then measure the number of connected components and the mean of the degree distribution for this sample, and perform statistical tests to see if the differences found are statistically significant.

For Sph1, the number of connected components observed was significantly smaller than the one observed for the random ensamble. For spheroids 3 and 4, however, the large CSC fraction made the experimental and ensamble values statistically undistinguishable from each other (with significance $\alpha_t = 0.001$). Nonetheless, the mean degree of the experimental networks was significantly higher than the one of the corresponding ensamble for all spheroids. This implies that stem cells are significantly more connected between them than it is expected if located randomly. This, in turn, reinforces our previous conclusion about the tendency of neighboring cells of the same stemness. Both of these results explain and characterize the *paths* formed by CSCs in the spheroid.

Jero
Para mi, algo de esto hay que mencionar en las conclusiones.

Conclusion

The method of image processing presented in the Methods Section is devoted to recognizing the location of the CSCs in microscopy images. Its originality is in the fact that we are not just looking at where the SOX2 fluorescence is high enough by

Field	Sph1, slice 3	Sph1, slice 4	Sph3, slice 3	Sph4, slice 2
Radius (μm)	113.25	99.07	102.04	87.58
Total Cells	183	252	240	112
Stem Cells	55	56	155	59
Differentiated Cells	128	196	85	53
Stem Cell Fraction	30.05%	22.22%	64.58%	52.68%
Homophily Ratio	0.6685 (↑)	0.7378 (↑)	0.7337 (↑)	0.7382 (↑)
Assortativity Coefficient	0.2210 (↑)	0.2634 (↑)	0.3937 (↑)	0.4716 (↑)
Stem Connected Components	5 (↓)	12 (↓)	2 (=)	2 (=)
Mean Degree (Stem Subgraph)	2.7273 (↑)	2.6429 (↑)	4.8258 (↑)	4.4746 (↑)

Table 1. *Summary of spheroids’ properties.* For each slice of analyzed spheroid, we report its approximate radius, the number of cells of each type it contains, the threshold in fluorescence determined by the clustering, indicating its robustness. Regarding the network’s homophily, we report the homophily ratio (fraction of edges of the Delaunay triangulation graph that link cells of the same stemness), the assortativity coefficient, and the number of connected components and mean degree of the stem subgraph. For these last four fields, we carried out statistical tests to see whether the obtained values were distinguishable from the ones we would obtain from a graph where stem cells were located randomly (but identical otherwise). We indicate that the value was higher/lower than in the random case, with a significance of $\alpha_t = 0.001$, by adding an arrow (↑ / ↓) next to the value. If the values are not significantly different, we add an equal sign (=).

a subjective criterion. The key is the use of the Gaussian Mixture Model to fit data and separate with a statistical criterion both cell populations. Having done this, the way of depicting the spheroid or what is done with the resulting data will depend on the questions that researchers have in mind. In our case, we presented an example that compares the experimental CSC distribution in tumorspheres with a previous computational model. Beyond the fact that we are pleased to find that *in silico* and *in vitro* experiments give similar outcomes, the method proposed here is straightforwardly applicable to similar experiments. Indeed, to definitively assess the CSCs distribution in a spheroid, we must be able to put into the microscope non-deformed spheroids and take pictures of several slices of them. Furthermore, according to our computational simulations, the confidence of our result will exponentially increase with the cultured time. Thus, we expect to perform or find longer-term experiments.

Finally, we would like to comment that full 3D simulations are being carried out. Our preliminary results still support the finding that CSCs are heterogeneously distributed inside the spheroid. However, the probability of finding CSCs on the border of the spheroid seems to be significantly enlarged.

Cerramos con chamuyo de Luciano sobre por qu’e es ’util esto en biolog’ia

Discutir biologicamente la ubicacion de las CSCs en los esferoides y el acceso a los farmacos

Methods

Topical subheadings are allowed. Authors must ensure that their Methods section includes adequate experimental and characterization data necessary for others in the field to reproduce their work.

Cell Culture and mammospheres assay

Human breast cancer-derived cell line MCF-7 was maintained in DMEM/F12 complete medium (DMEM F12 + FBS 10 + 1 Glutamine) at 37 C in a humified 5 CO2 atmosphere. The mammosphere assay was performed according to modifications[15] from the original protocol by[3]. Briefly, 6-well plates were treated with poly(2-hydroxyethyl methacrylate) to prevent cell adhesion. MCF-7 cells were seeded at 3000 cells/ml in complete mammospheres medium (DMEM/F12 + B27 2 + glutamine 1 +20 ng/ml EGF + 20 ng/ml bFGF). Cells were incubated at 37 C in a humified 5 CO2 atmosphere for 7-9 days, adding 0.5 ml of medium with growth factors every 48 hs. The resulting mammospheres were collected and attached by cytocentrifugation onto slides for immunofluorescent detection of Sox2.

Immunofluorescence

Immunofluorescent detection of Sox2 in mammospheres was performed as described in[15]. Briefly, slides were fixed in methanol and then washed 3 times in washed solution (PBS-BSA 0.1) during 5 min and permeabilized with PBS-BSA 0.1 + 0.3 triton X100 for 30 min at RT. Following permeabilization, the slides were incubated in blocking solution (PBS-BSA 0.1 + Normal Goat Serum) and the corresponding primary antibody (anti-Sox2, cat # PA1-16968, Thermo) ON at 4C. Next, the slides were washed 3 times in washing solution during 5 min and incubated with the corresponding secondary antibody (anti-rabbit 555, cat # A31572, Life Technologies).

Luciano’s explanation of the experimental methods, from the culture of the spheres, to the generation of the Z Stack − CTRL SOX2.czi files.

Data Analysis

Using the files obtained from the confocal microscopy, we wanted to segment the nuclei for identifying and counting the cells, and then associating the SOX2 fluorescence to the corresponding nuclei. As an example, we describe here the process for spheroid number 1 (Sph1 hereafter). The corresponding czi file contains data on three channels: an optic channel, a channel with the fluorescence of the nuclei, and another one with the SOX2 fluorescence. This data has three spatial dimensions (x,y,z), with 2292 pixels in both directions of the horizontal xy plane, and a thickness of 9 pixels in the z direction. The resolution of those pixels is the following

$$\Delta x = \Delta y = 1.2364633517553391 \times 10^{-7}\text{m} \simeq 0.12\mu\text{m}$$
$$\Delta z = 1.9999999999999999 \times 10^{-6}\text{m} \simeq 2\mu\text{m}.$$

Now, let us describe separately the processing for the SOX2 channel and the channel for the fluorescence of the nuclei.

The SOX2 Channel

The main goal for processing this channel was to get rid of the diffusing SOX2, revealing the spots where the marker had attached to the cells. To achieve this, we perform a reconstruction by dilation, a morphological image processing technique that allows us to obtain a denoised version of an image by subtracting a blurred version of it from the original one. We used Scikit −image’s implementation[16] of the morphology algorithms for this part (for further information, see scikit −image’s documentation). Fig. 2 shows this cleaning process for slice 3 (i.e., $z = 3$) of Sph1.

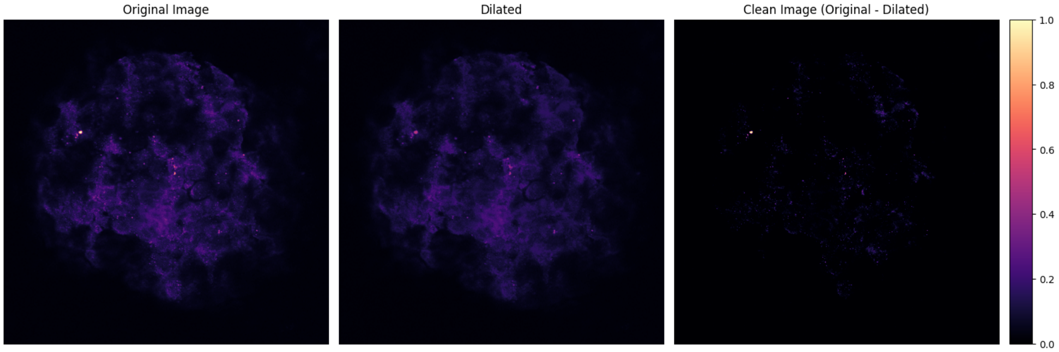


Figure 2. *Cleaning of the SOX2 channel.* The image shows, from left to right, the original image, a dilated version of the image (that can be thought of as a blurred version or background noise), and the cleaned image as the difference between the original and the dilated one. The original image corresponds to slice 3 (i.e., $z = 3$) of the SOX2 channel of Sph1.

The Nuclei Channel

For this channel, the main goal was using the nuclei to individualize and identify the cells. More specifically, we performed instance segmentation on the nuclei, extracting geometrical features such as the area and the coordinates of the center of the segmented objects. Let’s detail the procedure followed.

1. Contrast Limited Adaptive Histogram Equalization (CLAHE). This algorithm, also implemented in Scikit −image[16], is used for local contrast enhancement (see the documentation). This is for making edge recognition easier for the segmentation algorithm.

2. Morphological processing. For further improving the future edge recognition, we perform on the image a morphological opening (erosion followed by dilation), followed by an area closing (similar to a morphological closing, dilation followed by erosion, but using a deformable rather than a fixed footprint). The opening helps separate objects that may be in contact, while the area closing is used for removing small dark structures on the image. This later was used to avoid single nuclei being segmented as many objects due to the presence of dark spots within them. Once again, we use Scikit −image’s implementation[16] (see documentation).

3. Bilateral Denoising. We use an edge-preserving bilateral filter to denoise the image, averaging pixels based on their spatial closeness and radiometric similarity. This works both to reduce the noise introduced by the morphology operations, and by out-of-focus objects. See Scikit −image's documentation for more details.

4. Instance Segmentation. For identifying the nuclei, we use a pre-trained Stardist [17] model. More specifically, we use the 2 D_versatile_fluo model (see documentation). This is a convolutional neural network with a U-Net architecture, trained on fluorescence microscopy images (similar to the ones on our experiment) to identify star-convex polygons bounding the nuclei. We chose this model instead of instance segmentation with bounding boxes because in this way we do not need a subsequent shape refinement. Furthermore, semantic (per-pixel) cell segmentation requires a subsequent pixel grouping that can result in segmentation errors such as falsely merging bordering cells. Such errors would be very likely to happen in our case, since the images portray situations of very crowded cells. Star-convex polygons provide a much better shape representation that overcomes these difficulties.

The complete process can be seen in Fig. 3. From the segmentation, we extract many geometrical features, their centers in particular.

Tessellation for cell region identification

Since the SOX2 marker does not necessarily bound to the nucleus of the cell, but rather to its cytoplasm, we need a way of assigning each point in space (i.e., each pixel) to a single cell. That would enable us to associate the fluorescence of the marker with a given cell. For doing this, we approximated the division of the space associated to each cell by a Voronoi tessellation. We use the centers of the segmented objects, filtering out the ones corresponding to cells that do not belong to the spheroid (if needed). The result can be seen in Fig. 4a. It is worth mentioning that we have used a number of artificial points to limit the regions on the edge of the sphere. Otherwise, they would be unbounded regions, which has proven to be very inconvenient (besides being an inappropriate representation of the cells).

With this, we can now associate SOX2 fluorescence inside a Voronoi region to the cell of that region. More concretely, we associate the cell to the sum of the fluorescence intensity of each pixel of the SOX2 channel in that region. This association can be seen in Fig. 4b. In order to better visualize the tessellation, we overlay it on the nuclei fluorescence and the cleaned SOX2 channel in Fig. 4c.

Stemness threshold in SOX2 fluorescence intensity

esto lo pongo como los m'as relevant del m'etodo en discussion. Hay que maquillarlo en consecuencia In order to identify which of these cells are stem, we need to define a threshold in the total SOX2 fluorescence of the corresponding cell region that divides non-stem from stem cells. This can be done by using a clustering algorithm to group the intensities of the cells. It's worth mentioning that an "elbow" plot using k-means clustering suggests that 2 is an appropriate number of clusters for this data. In what follows, however, we will consider a fixed number of two groups. Data points will be assigned to each group according to a Gaussian Mixture Model (GMM) fitted to the data. In short, this method consists in assuming that the points are generated by two normal distributions, and fits their means and standard deviations using the maximum likelihood criterion. We used Scikit −learn's implementation[18]; for additional details, see the documentation.

A necessary consideration to be made is that, since the fitting of the distributions is sensitive to extreme values, outliers in each direction (5th and 95th percentiles) will be directly assigned to the corresponding category (non-stem for the lowest values and stem for the highest), and will not be taken into account when fitting the GMM. Continuing with our example, we plot the histogram of SOX2 fluorescence intensities for Sph1 in Fig. 5. The threshold value V turns out to be of approximately 71.4. This implies that $S_V = 55$ out of the total $N = 183$ cells in the spheroid are stem cells, which represent a fraction of $f_V \simeq 0.3005$ of the total.

Something worth mentioning is that, for every case, we've checked that the random seed of the clustering algorithm didn't modify the threshold significantly. Usually, there are a couple of values to which the threshold converges. Checking for robustness of the clustering means that these values are close to each other, regardless of the random state used and the clustering algorithm. If these values vary a lot or are not close to each other, the clustering becomes unreliable. For the case of our example, both GMM and K-means yield a threshold of 71.4, for a big number of random seeds.

A more intuitive way of visualizing this result is taking Fig. 4c and coloring each region according to its clustering. This can be seen in Fig. 4d where regions corresponding to stem cells have been colored red, and the non-stem ones blue.

For Sph1, there were two layers of cells that laid one on top of the other. The layer on top is the one we've seen in slice 3. The layer on the bottom could be clearly seen in slice 4. Fig. **??** shows the end result for slice 4, of the process described for the previous slice. For this bottom slice, 56 out of 252 cells turned out to be stem cells (which represent a 22.22%), with the intensity threshold located at approximately 110.00. With this, the final result we get it's just the tessellation with the clustering indicated; we no longer show the fluorescence channels. This can be seen, for instance, in Fig. **??**, and Fig. **??**.

Jero

Broken ref. cause sb deleted the im-

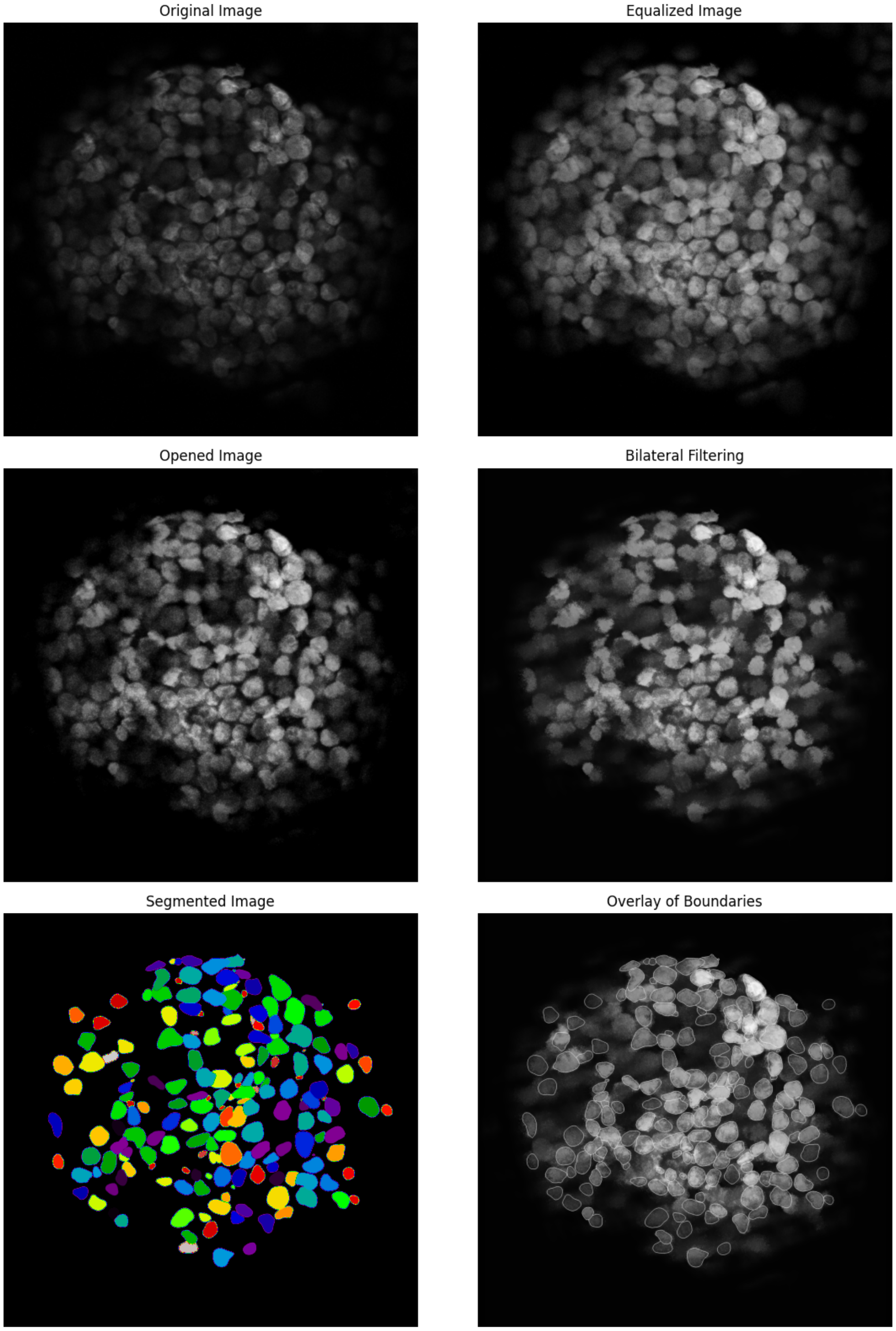


Figure 3. Ponerlas 3 arriba y 3 abajo*Processing of the nuclei fluorescence channel.* The images (from left to right and top to bottom) show the different steps in the processing of an image. The one in the bottom right corner shows the borders of the identified structures over the image provided to the model. The darker, unrecognized nuclei correspond to cells that are outside (in this case behind) the focal plane.

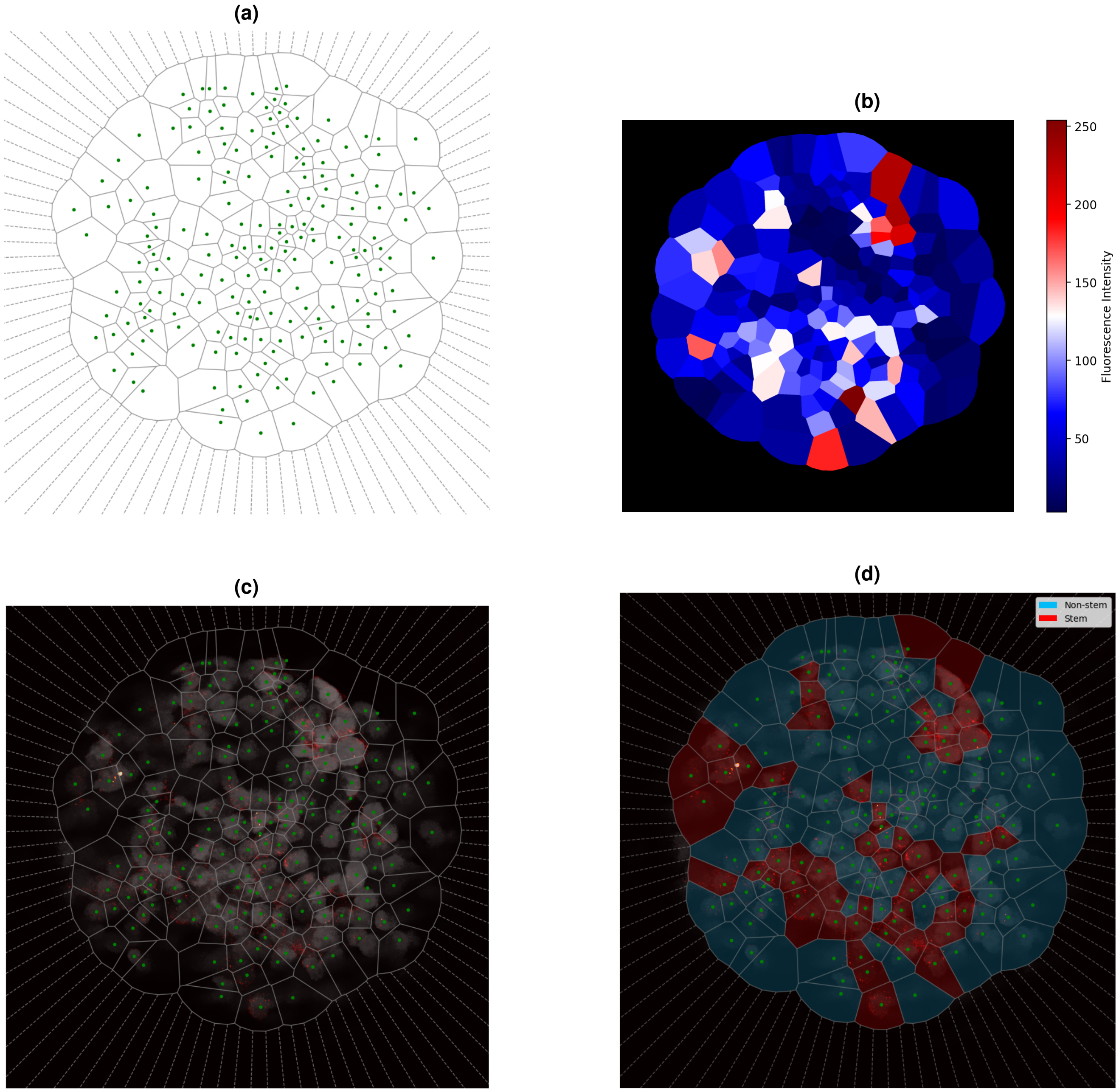
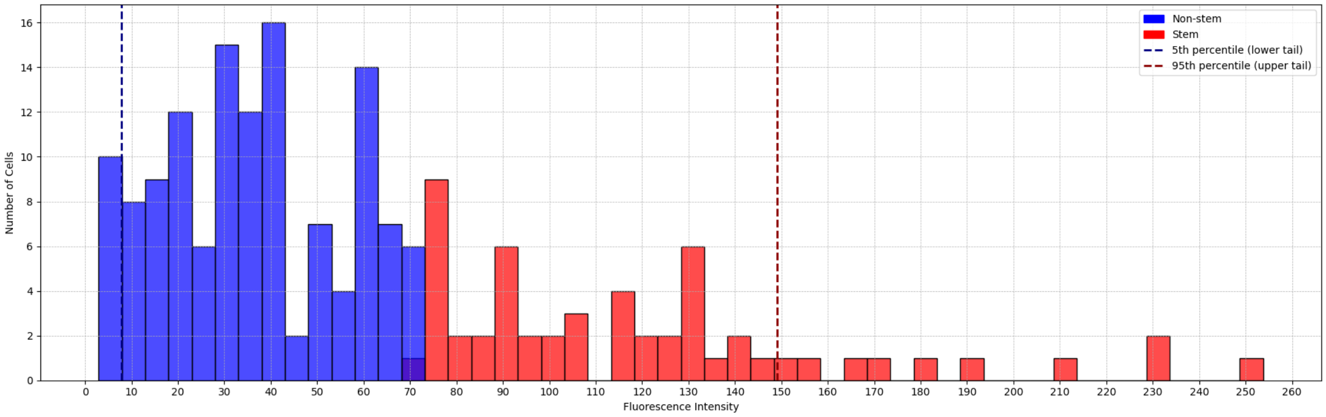


Figure 4. un solo archivo . d) cambiar non-stemFrom Voronoi Tessellation to CSC recognition. **(a)** Voronoi tesselation. The green dots, corresponding to the centroids of the cells, are used to estimate the cell area. **(b)** Regions colored according to SOX2 fluorescence intensity. For each region of the tessellation, the sum of its SOX2 fluorescence intensity is computed. **(c)** Tessellation overlaid with nuclei, grey scale, and SOX2, red palette, channels. The centroids are shown as green dots, and the boundaries between Voronoi regions are plotted with gray lines. **(d)** Tesselated regions colored in RED for Cancer Stem Cells and BLUE for Differentiated Cells according to SOX2 fluorescence content. If the fluorescence surpasses the threshold given by the GMM, the cell is considered a stem one.

Possible variations of the method

The details when implementing this method depend on the set of images obtained from the microscopy. In the case of Sph1, there were two layers of cells, and only two slices were able to properly show each of them. The images for the other slices were out of focus. However, had they been on focus, we could have tried to take advantage of the extra information provided by



agrandar etiquetas . cambiar non/stem

Figure 5. *Histogram of SOX2 fluorescence intensities, colored by the GMM clustering.* Intensity values were clustered into non-stem and stem ones by fitting a GMM. Regions with intensities in the (5th) 95th percentile are automatically considered (non-) stem ones and are not taken into account when fitting the GMM.

this other slices. This can be done in different ways, with varying degrees of effort.

As an example, assume that the same layer of cells is seen in-focus by more than one slice. Then, the easiest approach would be to associate the Voronoi regions of the same cells throughout the slices. This could be done by setting two regions (on different slices) as belonging to the same cell if the distance between their centers is below a given threshold. A more sophisticated approach would be to do the 3D segmentation of the nuclei, and then obtain the 3D Voronoi tessellation. This last approach carries with it the drawback of having to label training and testing data, so it is to be avoided unless necessary.

The benefit of both these types of 3D tesselation, is that we can take advantage of the SOX2 fluorescence through the slices for a better clustering. Also, if one doesn't have clearly defined layers that don't share any cells, a 3D tessellation could be the only way to reliably count and cluster cells.

Another situation that one could encounter, would be having many isolated cells surrounding the spheroid, that do not belong to it. Filtering them out automatically is not hard. The process would be the same, we segment all the objects in the image and use all of them for the tessellation. We add the SOX2 fluorescence intensity of each region, but before clustering, we filter out the regions whose center is are further away from the center than a certain threshold. This was done for slice 4 of Sph1, but the effect was very subtle in that case because the cluture's environment was very clean. A better example is the one of slice 3 of Sph3, shown in Fig. **??**. Here, we calculated the center of the spheroid x_S as the mean position of overall segmented objects. Then, we defined the maximum distance to this center d_{max}, as the maximum distance between the centers of the segmented objects and the spheroid's center x_S. The threshold for a region to be included in the clustering was set to αd_{max}. For this particular case, a value of $\alpha = 0.55$ results adequate. There were other cases, namely the one of slice 2 of Sph4, whose edges were even more disaggregated, which led us to choose a smaller (tighter) threshold of $\alpha = 0.5$ (see Fig. **??**). The reason for excluding isolated cells from the clustering is that the cleaning of the SOX2 channel doesn't attenuate the fluorescence for the regions corresponding to these cells. This happens because the cleaning process is designed to eliminate noise; as isolated cells are surrounded by the black background, any SOX2 fluorescence around them will not be considered noise. For this reason, they are mostly clustered as stems, even though they didn't show high fluorescence in the original image.

CSC connectedness and path forming

From seeing the spatial location of the stem cells in the different spheroids, it's intuitively clear that their distribution is far from random. They seem to be primarily forming clusters, therefore breaking the homogeneity of their distribution, something that would be expected if located randomly. Here, we'll develop a quantitative characterization that justifies this intuition.

The first step in our analysis is to construct the Delaunay triangulation of the Voronoi tessellation, which corresponds to its dual lattice. This is nothing else than the graph given by using the centroids of the Voronoi as nodes, and adding a link between two nodes if the corresponding cells in the Voronoi diagram are neighboring cells. One such graph can be seen for our example of Sph1, slice3, in Fig 6. Then, we want to see if the graph presents network homophily with respect to the stemness attribute. Presenting homophily by stemness means that nodes in the network have preferential attatchment for other nodes of the same type. That is, cells of the same stemness are more likely to be connected than cells of different stemness.

We characterize network homophily by looking at the following parameters:

1. *Assortativity coefficient.* It is the most direct measure of the presence of preferential attachment. It equals 1 for perfect

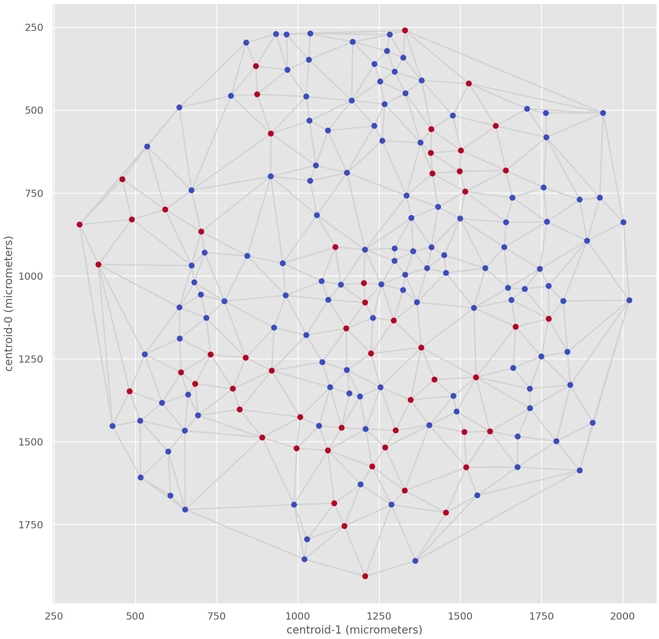


Figure 6. eliminar etiquetas de los ejes, lines de grilla y fondo gris. Agrandar el tama;o de los puntos.*Delaunay triangulation (slice 3, Sph1).* The network shows Voronoi regions as nodes, and adjacency between regions as edges. Regions corresponding to stem cells are plotted as red dots, while differentiated cells are plotted as blue dots. The axis represent spatial coordinates in the focal plane, measured in micrometers.

assortativity (perfect preferential attachment), 0 for non-assortative networks (i.e., no preferential attachment), and -1 for perfectly disassortative networks (perfect heterophily).

2. *Homophily ratio.* Is the fraction of the edges of the graph that link nodes of the same stemness.

3. *Number of stem connected components.* Looking at the subgraph formed by considering only the stem cells of the original graph and the links between them (stem subgraph), we calculate the number of connected components. If stem cells form clusters, the number of connected components will be smaller than expected for a random distribution.

4. *Degree distribution of the stem subgraph.* We can calculate the stem subgraph and its degree distribution. If, for instance, the mean degree is higher than expected for randomly distributed stem cells, it means that stem cells are more connected between them than what would be expected from the random case.

Now, these parameters alone do not tell us much, we need to calculate their expected values in the case where stem cells are randomly located. The procedure for doing this is the following. We take the graph of the network in question, and we randomly redistribute the location of the stem cells, maintaining its number and everything else in the network. Then we calculate the parameters previously mentioned, and repeat the process for a number of times to have a big enough sample (ensamble). Here, we have used a sample of size 10000 for each case. This allows us to calculate confidence intervals for the random parameters, and do statistical tests (z-test) for deciding whether a given value could be obtained from the distribution of the sample (i.e., if the parameter is significantly different from the random case). In every case, the significance α_t used was 0.001, but results are robust against reducing α_t.

In the case of Sph1, both for slices 3 and 4, we find that the assortativity coefficient, the homophily ratio, and the mean of the degree distribution for the stem subgraph, are all significantly higher than for the random case, while the number of connected components was significantly smaller. The same happens for Sph3 and Sph4, except that the number of stem-connected components isn't smaller than expected from the random case due to the high number of stem cells.

References

1. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc.Nat.Ac.Sc.* **100**, 3983–3988 (2003).

2. M, K. & MS, W. Implications of the cancer stem-cell hypothesis for breast cancer prevention and therapy. *J Clin Oncol* **26**, 2813–2820, DOI: 10.1200/JCO.2008.16.3931 (2008).

3. Dontu, G. *et al.* In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* **17**, 1253–1270, DOI: 10.1101/gad.1061803 (2003).

4. Ehmsen, S. *et al.* Increased cholesterol biosynthesis is a key characteristic of breast cancer stem cells influencing patient outcome. *Cell reports* **27**, 3927–3938 (2019).

5. Pastrana E, D. F., Silva-Vargas V. Eyes wide open: a critical review of sphere-formation as an assay for stem cells. *Cell Stem Cell* **8**, 486–498, DOI: 10.1016/j.stem.2011.04.007. (2011).

6. Kessel SL, C. L. A high-throughput image cytometry method for the formation, morphometric, and viability analysis of drug-treated mammospheres. *SLAS Discov* **25**, 723–733, DOI: 10.1177/2472555220922817 (2020).

7. Vieira LC, V. D., Costa RS. Fractional calculus as modelling tool. *Fractal Fract.* **7**, 595, DOI: 10.3390/fractalfract7080595 (2023).

8. Benítez, L., Barberis, L. & Condat, C. A. Modeling tumorspheres reveals cancer stem cell niche building and plasticity. *Phys. A: Stat. Mech. its Appl.* **533**, 121906, DOI: 10.1016/j.physa.2019.121906 (2019).

9. Barberis, L. M., Benitez, L. & Condat, C. Elucidating the Role Played by Cancer Stem Cells in Cancer Growth. *MMSB* **1**, 48–54 (2021).

10. Benítez, L., Barberis, L., Vellón, L. & Condat, C. A. Understanding the influence of substrate when growing tumorspheres. *BMC Cancer* **21**, 276, DOI: 10.1186/s12885-021-07918-1 (2021).

11. Fotinós, J., Barberis, L. & Condat, C. Effects of a differentiating therapy on cancer-stem-cell-driven tumors. *J. Theor. Biol.* **572**, 111563, DOI: 10.1016/j.jtbi.2023.111563 (2023).

12. Barberis, L. Radial percolation reveals that Cancer Stem Cells are trapped in the core of colonies. *Pap. Phys.* **13**, 130002–130002, DOI: 10.4279/pip.130002 (2021).

13. Chen, Y. C. *et al.* High-throughput single-cell derived sphere formation for cancer stem-like cell identification and analysis. *Sci. Rep.* **6**, 1–12, DOI: 10.1038/srep27301 (2016).

14. Wang, J. *et al.* A novel method to limit breast cancer stem cells in states of quiescence, proliferation or differentiation: Use of gel stress in combination with stem cell growth factors. *Oncol. Lett.* **12**, 1355–1360, DOI: 10.3892/ol.2016.4757 (2016).

15. Leis, O. *et al.* Sox2 expression in breast tumours and activation in breast cancer stem cells. *Oncogene* **31**, 1354–1365 (2012).

16. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453, DOI: 10.7717/peerj.453 (2014).

17. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, 265–273 (Springer, 2018).

18. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

The authors thank Dr. CA Condat for insightful discussions and comments. The confocal imaging was performed by CPA del MICROSCOPIO.

Funding

The theoretical work was supported by SECyT-UNC (Project 113/17) and CONICET (PIP 11220110100794), Argentina. Experiments were supported by FINANCIAMINEOT LUCIANO.

Author contributions statement

LB and LV conceived the experiment, LV conducted the experiment, JF conceived and performed the image processing. All authors analyzed the results and, wrote and reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** ALl data and codes are accessible under reasonable request. **Competing interests** The authors declare no competing interests.

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.