**Title:** Automated Single-Cell Quantification of Breast Cancer Marker Evolution from Serial Immunofluorescence Assays

**Authors and Affiliations:**

Rikhil Rao1

1 Canyon Crest Academy

# **Abstract**

Despite advances in breast cancer diagnoses and treatment, it continues to be the second leading cause of cancer death in women. Recent interest in managing and prognosticating breast cancer has been centered around examining single-cell and spatial expression patterns on molecular imaging, such as immunofluorescence and immunohistochemistry ((Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).)). However, current techniques for this single-cell quantification are disjointed and steps are not formatted to be compatible.

The purpose of this project is to develop a simple, easy to use, end-to-end pipeline to quantify how much a patient’s cancer changes at different timepoints in their treatment. To accomplish this, serial cyclic immunofluorescence images acquired over the course of a patient’s breast cancer treatment were gathered from the Human Tumor Atlas Network (HTAN). Two images of different samples were chosen from time point 0 and one image from time point 1 was chosen. Each image consisted of 25 unique channels, including DAPI, HER2, and ER. These images were used to train a patient-specific Variational Auto-Encoder (VAE), which was then used to quantify how much the cancer changed. In order to train the model, individual cells were segmented using an open-source tool called MCMICRO, resulting in around 14,000 cells in each image.

By running these analyses, the tumor phenotype of the patient (e.g., HER2 negative, estrogen receptor [ER] positive) was able to be automatically determined. In addition to this the MCMICRO quantification outputs show that the marker expression changed at a later time point as the patient underwent hormone therapy for their breast cancer. For example, the ER expression decreased 2-fold and the Ki67 expression increased 12-fold. Additionally, the VAE latent space plots show significantly different distributions at different time points, implying that the network was able to encode the baseline single-cell expression patterns.

This research shows that automated single-cell quantification of marker expression in breast cancer can be important to track tumor evolution over time, which can be quantified, visualized, and predicted using a VAE.

## **Introduction**

Cancer has become one of the main threats to human health and life. Among all types of cancer, breast cancer is very common among women and diverse in character. About 70 percent of all breast cancers in women have overexpression of receptors (proteins) that attach to estrogen or progesterone ((Breast Cancer Hormone Receptor Status | Estrogen Receptor. https://www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html)). Breast cancers without ER (Estrogen Receptor) and PR (Progesterone Receptor) are known as HR (Hormone Receptor)-negative. Estrogen receptor and progesterone receptor tests are used to help guide breast cancer treatment with hormonal chemotherapies such as tamoxifen ((Estrogen Receptor, Progesterone Receptor Tests: MedlinePlus Medical Test. https://medlineplus.gov/lab-tests/estrogen-receptor-progesterone-receptor-tests/)). However, patients with identical hormone receptor phenotypes can have variable responses to therapy ((Gupta, R. K. & Kuznicki, J. Biological and Medical Importance of Cellular Heterogeneity Deciphered by Single-Cell RNA Sequencing. *Cells* **9**, 1751 (2020).)). Recent research shows that cancer cells and cells within the tumor microenvironment can have variable phenotypes across parts of the tumor, and more specific analysis of these distribution patterns can help determine disease progression and response to treatment ((Li, A. *et al.* Characterizing advanced breast cancer heterogeneity and treatment resistance through serial biopsies and comprehensive analytics. *Npj Precis. Oncol.* **5**, 1–12 (2021).)).

Immunofluorescence is a powerful approach for getting more information about cellular structures and biomarker expressions. Augment simple immunofluorescence, cyclic immunofluorescence (CyCIF) is a novel technique to study expressions and locations of a multitude of specific biomarkers in the same sample ((Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y. & Sorger, P. K. Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264 (2016).)). Since the extent of cellular heterogeneity in breast cancer has been hypothesized to impact diagnosis, response to therapy, and long-term outcome, CyCIF of breast cancer biopsies has potential to guide breast cancer management in the future.

CyCIF generates gigabytes of image data that scientists and technicians do not have the bandwidth to process or ability to identify specific patterns. In contrast, artificial Intelligence based computational pathology has shown increased diagnostic accuracy and reduced turnaround times ((Basu, K., Sinha, R., Ong, A. & Basu, T. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? *Indian J. Dermatol.* **65**, 365–370 (2020).)). Computer aided image analysis of CyCIF data has the potential to make this complex information more accessible in the diagnostic workflow of breast cancer ((Cheung, A. M.-Y. *et al.* Quantitative single-cell analysis of immunofluorescence protein multiplex images illustrates biomarker spatial heterogeneity within breast cancer subtypes. *Breast Cancer Res.* **23**, 114 (2021).)). In this study, we show the feasibility of a fully automated deep learning-based pipeline for analysis and tracking of biomarker expression patterns in serial breast cancer biopsies.

## **Results**

The result is a pipeline that takes in a set of molecular imaging files (often 10-100 GB in size) and outputs the single cell quantification for each channel of the imaging.

### VAE Training

In the training dataset, the pipeline took an input of 30 channels (DAPI, HER, ER, PCNA, Ki67, etc.) and accurately segmented ~16,000 cells from which cell-specific marker expression quantifications were obtained. Execution of the pipeline on the testing datasets at the same and later time points resulted in ~21,000 and ~4000 identified and quantified cells, respectively. Figure 1 shows an exam cropped section of one of the biopsies run through the MCMICRO UNMICST cell segmentation algorithm.

A collage of different images

Description automatically generated

Figure 1: Example segmentation for one region of interest (ROI) from biopsy. Passing the image on the left through the MCMICRO pipeline generates the images on the right. The images on the right are the masks of the cells that were in the ROI.

### Marker Quantification Results

Looking at the quantification of the markers at different timepoints allows us to make claims about the cancer progression. For example, the increase in the expression of Ki67 indicates to us that the cancer is proliferating at a much higher rate.**A graph of different colored bars

Description automatically generated**

Figure 2: The concentration of several markers in various images at different timepoints

### VAE Training and Analysis

The VAE model was trained for 200 epochs and reached a test loss of 0.1159 and a validation loss of 0.1164. The training and testing loss is reflected in the quality of the image generated as seen in Figure 4. Figure 4 shows an image fed into the VAE and the rebuilt version of that image from the VAE. The drop in quality is because of the lack of latent vectors used to represent the 24x24 images but is necessary to interpreting the data in a friendly way. Figure 5 provides a visual representation of the latent vector at different timepoints and shows the difference between the timepoints. From the graphs a clear distinction between each timepoint can be compared using the KL Divergence Statistic to quantify the difference in between two timepoints.

A graph of a train test

Description automatically generated

Figure 3: The above image shows the model loss of our trained VAE. Model loss = latent loss + reconstruction loss. Reconstruction loss is the loss that is experienced when rebuilding the image and latent loss is the loss that is experienced by the latent space.

A screenshot of a computer generated image

Description automatically generated

Rebuilt image by VAE

Image fed into VAE

Figure 4: The quality of the rebuilt image shows the disadvantage to only using 2 latent vectors to represent the entire 24x24 image.

## A picture containing graphical user interface Description automatically generated

Figure 5: The larger KL Divergences for the comparisons *between* timepoints than those *within* a timepoint suggests that the VAE is learning the patient’s baseline tumor phenotype.

## **Discussion**

This work shows the feasibility of using an automated pipeline to create a personalized VAE that can detect tumor divergence during a patient’s breast cancer treatment course. Future work on this topic should focus on better examining the meaning of the latent space vectors of the VAE and how they relate to patterns across image channels. Many of the shortcomings of this project have to do with sacrificing the quality of the results for interpretability of the results. For example, adding another dimension to the latent vector would have increased the accuracy of the regenerated image but would make it much harder to quantify the difference between two timepoints as the data would be 3 dimensional. Along with that, this project is only able to quantify how much the cancer changed and not if it changed for worse or for better, though this information can be interpreted from the biomarker expression plots. With more analysis of the latent space combined with other analyses performed on the biopsy, it would be possible to better understand how much the cancer changed and why the cancer changed.

## **Methods**

### Data:

The Human Tumor Atlas Network (HTAN) open dataset of imaging was accessed on their website. One breast cancer patient with serial biopsies each with CyCIF images available was selected. Imaging data of three biopsies (two at a concurrent time-point and one at a follow-up time-point) for this patient was gathered. Raw CyCIF files were downloaded via the Google cloud command line interface (gCloud CLI). All the files were downloaded in DICOM format, the common file type for medical images. Each DICOM file was a slide of the complete biopsy and represented a different marker.

To run the data through the MCMICRO pipeline, a csv file that contains information on the marker associated with each channel in the DICOM image was necessary. Each DICOM file that was downloaded was associated with a singular marker, so the CSV file was needed to compile all the different markers that were present in the biopsy. Headers for all the DICOM files that were downloaded were parsed and used to generate a csv file for each biopsy.

### MCMicro:

Using nextcloud, the MCMICRO pipeline was run on the dataset with the generated csv file.

To create training data for the VAE, the segmentation of the dataset that was performed by the MCMICRO pipeline was used to generate small images (24 by 24 pixels) for each of the cells that were found. The image size is 24 x 24 because it was the maximum size of any given cell that was segmented so it would for sure capture each of the cells that were segmented. The data was then saved into a numpy array.

### VAE:

The Variational Auto-Encoder was created using Keras, a neural network library. The model that was created has 4 2D convolutional layers, 2 dense layers, 1 flatten layer, and 1 2D Convolutional up sampling layer. To train the model, saved data was loaded in the form of a numpy array and then partitioned so that 75% of the data stored in the file was used for training while the other 25% was used for testing.

The VAE was trained on one biopsy from one time point and tested on held out test data from the same biopsy/time point, a different biopsy at the same time point, and a different biopsy at a later time point, The latent vectors of each run were visualized and used to quantify how much the cancer changed. The KL divergence statistic was used to quantify the difference in the probability distributions.

## **Conclusion**

This project applied a Deep Learning-based segmentation module to quantify tumor marker expression in breast cancer molecular imaging over time as a patient underwent breast cancer treatment. A VAE trained on baseline data showed that it was able to track and quantify the differences in single-cell expression over time. Future research on this topic should focus on validating this pipeline in multiple other subjects who have serial biopsies and better understanding the implications in tumor progression from the latent space changes.

## **Acknowledgments**

I would like to thank our computer science teacher and guide Mr. Kevin Hare for the support and encouragement to complete this project. I would like to acknowledge the guidance and help received from Mr. Gunvant Chaudhari, who taught me cancer biology and machine learning concepts that helped me to implement this project. Finally, I would like to thank my parents for being a source of information and providing the components, supplies needed for this project and to proofread the draft write-ups and material to evolve to the current version.