Image Analysis and Platform Development for Automated Phenotyping in Cytomics

Yan, K.

Chapter 1

Introduction to High-throughput Cytomics

1.1. Bioinformatics and Cell Biology

Bioinformatics is the research field that attempts to extract comprehensive information from large quantities of biological data via an integration of methods from computer science, mathematics, and biology [1]. It is an interdisciplinary study of automation for the retrieval, storage and analysis of biological data. The term bioinformatics, so it is posed, was first in 1970 by two Dutch complex system researchers, Ben Hesper and Paulien Hogeweg [2]. Since then, bioinformatics has been the key to functional genetics and system biology experiment in order to understand mechanisms behind cell behavior and cell system. In current cell biology research, bioinformatics plays a crucial role in the correlation modeling between cell **genotype** and **phenotype**.

In cell biology, the term **genotype** refers to the unique genetic hereditary and expression pattern while the term **phenotype** refers to observable cell properties including morphology and migration behavior. The correlation modeling of **genotype-to-phenotype** is a study of the correlation/causality between characteristic patterns of gene expression and recognizable cell properties. The genotype-to-phenotype correlation is the foundation of a systematic understanding of molecular architecture and control mechanisms behind different cell behavior. This correlation is of particular importance to cell behavior related diseases. However, due to the amount of information to be analyzed, the study of genotype-to-phenotype correlation must be done in an automated fashion.

To that end, modern developments in quantitative microscopy and laboratory robotics have provided the necessary hardware to the automated study of genotype-to-phenotype correlation. These studies are often referred as the –omics studies[3]. Depending on the study subject and biology of interest, omics studies employed by molecular genetics and cell biology branched out to a number of fields including genomics, **transcriptomics**, proteomics, metabolomics and **cytomics**.

Transcriptomics is the study dedicated to the quantitative analysis of RNA expression. In a living cell, RNA is produced by transcripting DNA strains. The expression pattern of RNA is considered an indication of protein production related to cell behavior. The study of the correlation between RNA expression and cell behavior is an essential part of genotype-to-phenotype study. Much of it is accomplished via the application of an high-throughput screen technology known as the **microarray**. The microarray technology is a method originally derived from Southern blotting [4] developed by Sir Edwin Mellor Southern in 1992. Essentially, microarray technology relies on base-pairing and hybridization of RNA strains to quantify the expression level of targeted RNA strains.

Transcriptomics focuses on the quantification of the genotype, whereas **cytomics** is dedicated to the quantitative analysis of the cell phenotype using high-throughput screen technology; this is known as the high-throughput/high-content cell screen or simply **HT/HC screen**. The HT/HC screen is achieved via the introduction of automated microscopy. Modern HT/HC screen analysis produces data on terabyte scale. Such a data volume poses great difficulty in manual analysis in early cytomics studies. To that end, computer science and information theory including image analysis, pattern recognition and machine learning have been included

to provide automated solution for image data processing. Due to the unique characteristics of biological image data, image analysis solutions need to be tailored. Therefore, we systematically investigate the limitations of generic image analysis solutions in HT/HC screen experiments; and based on the limitation studies, this thesis will focus on the design of dedicated image analysis solutions for HT/HC screen studies.

1.2. Image Analysis in High-throughput/High-Content Screen

At the current stage, cytomic experiments frequently employs either **flow cytometry** or **high-content screen** techniques to capture cell phenotypes.

Flow cytometry (cf. Figure 1-1) is a laser based technology employed in cell counting, sorting [5][6], and biomarker detection by suspending cells in a fluid stream and passing them through an electronic detection apparatus.

The **high-throughput/high-content screen** (HT/HC screen) [7] is a microcopy based method aiming for the visualization and capture of cell phenotypes over a large range of experimental settings (cf. Figure 1-2 & Figure 1-4). With different microscopy techniques, the HT/HC screen can serve different biological studies.

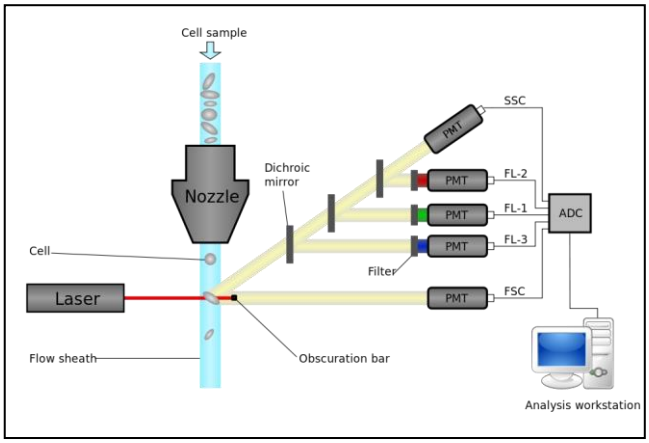

Figure 1-1 the workflow of a multi-color flow cytometry

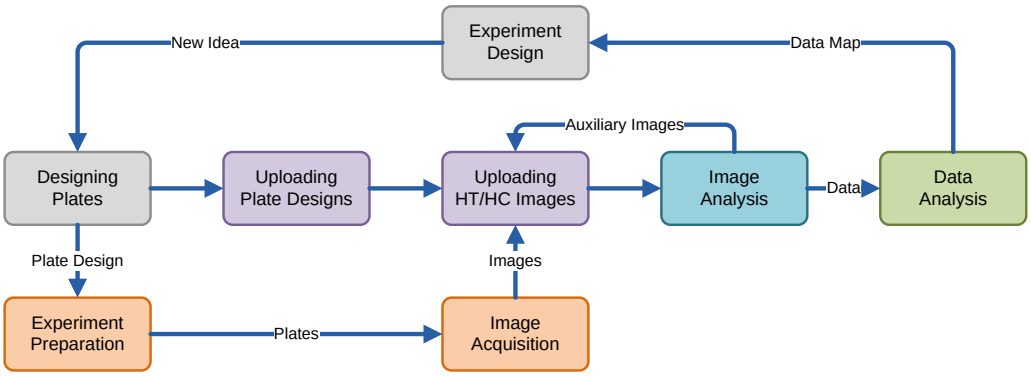

Figure 1-2 the general workflow of HT/HC screen study

In cytomics studies, the flow cytometry technology is frequently used to sort and quantify different types of cells [5][6][8] due to its fast analysis and high sorting accuracy. However, the flow cytometry does not provide the possible to study live cell behavior. Compared to flow cytometry, HT/HC screen using automated microscopy can capture both cell morphology and cell migration, i.e., size, polarization and migration speed. The employment of an HT/HC screen allows the extraction of phenotypical quantification; combined with microarray technology it

allows further correlation modeling of genotype-to-phenotype can be accomplished[9][10][11][12][13][14].

An HT/HC screen study consists of a sequence of five steps[9][14]: (1) **experiment design**, (2) **experiment preparation**, (3) **image acquisition**, (4) **image analysis**, and (5) **data analysis**. In each step, the raw data are further transformed into comprehensive data representation to support validation of the initial hypothesis.

1. Experiment Design

The experiment design is the first step in a HT/HC screen. It is the step to draw the blueprint for the HT/HC screen study to serve the research question. During the experiment design, the researchers must answer two essential questions: (1) what information must be captured and (2) how to capture the information. It is very difficult to be precise on how these two questions should be answered. Instead, we will give two examples of experiment design to explain the procedure.

Example 1: Growth factor regulated cancer metastasis [10][15] (cf. Figure 1-6a & Ch. 4)

In the experiment design of this HT/HC study, the researchers must first select a siRNA library which is potentially related to cell migration. Second, they must choose a cancer cell line demonstrating strong phenotypical plasticity, thus allowing to capture more observable differences in cell phenotype. Third, they must design a layout (cf. Figure 1-3) to fit the treated cells into a culture plate (cf. Figure 1-5). At the same time it is decided which microscopy technique is most suitable to capture the cell phenotype.

Example 2: Study of matrix adhesion dynamics [16] (cf. Figure 1-6b & Ch. 5)

In the experiment design of this HT/HC study, the researchers must first decide what combination of microscopy techniques can be used to capture matrix adhesion (subcellular structure) and cell migration (cf. Figure 1-4). Finally, they must decide how to quantify the relationship between the dynamics of matrix adhesions and cell migration.

From these two examples it is clear that for different biological questions, the experiment design step may consist of different tasks. As a result, there are very few limitations to the step of experiment design which makes it vulnerable to user generated error. In Chapter 6, we will demonstrate a flexible computer-aided experiment design interface for HT/HC screen studies.

DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
pIGFP				GB1			

Figure 1-3 plate layout of a 8x12 culture plate used in RCM3 growth factor regulation (cf. chapter 4) in Excel spreadsheet

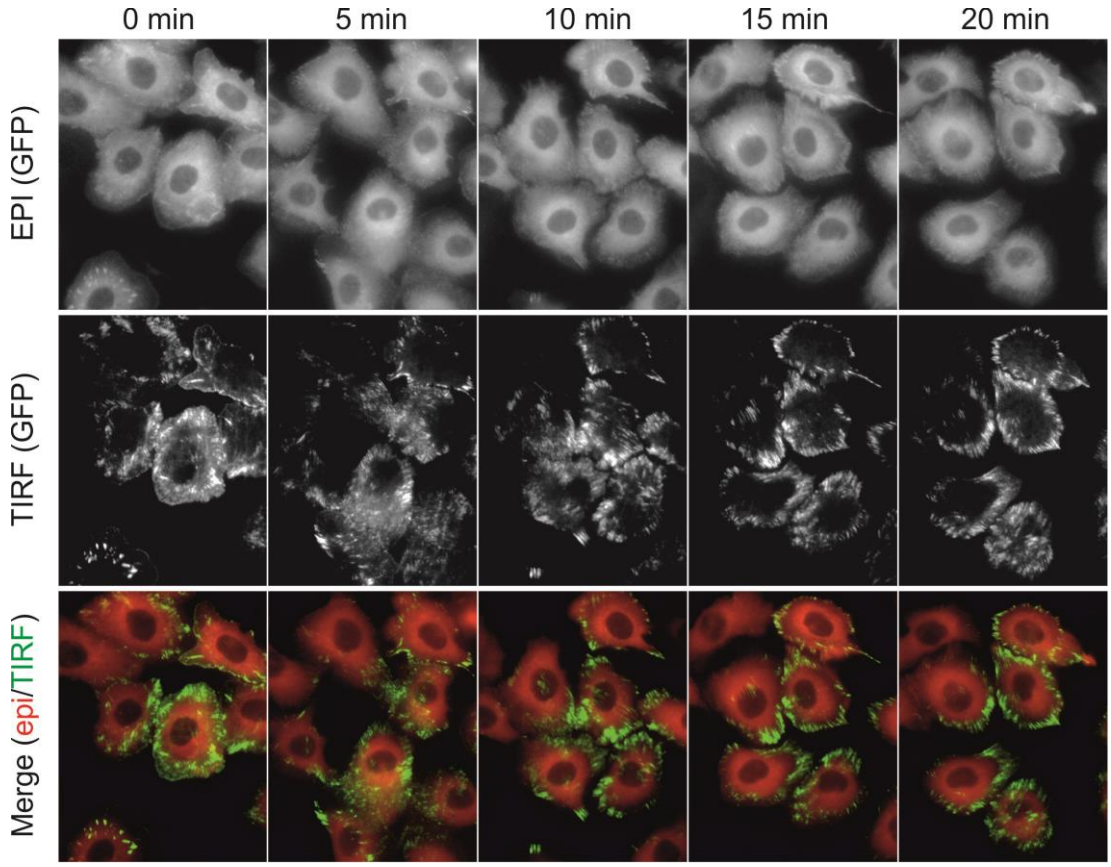

Figure 1-4 sample frames from a temporal-spatial 2D imaging setting consists of a TIRF channel and a epifluorescence channel

2. Experiment Preparation

During the experiment preparation, the researcher will conduct the experiment following the initial experiment design. Specimens, usually *in vitro* cell lines, are prepared and put into a culture plate (cf. Figure 1-5). However, the cell behavior is often subjected to environmental fluctuations such as CO_2-level, temperature [17][18], or cell-to-cell variability [19]. As a result, it is likely that an individual specimen may not homogeneously express an expected phenotypical signature. These typical wet-lab related problems are, however, beyond the scope of this thesis.

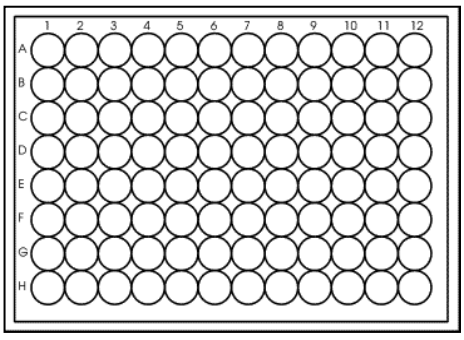

Figure 1-5 layout of a 96 well culture plate

3. Image Acquisition

In the image acquisition step, according to the plate layout, images will be produced (cf. Figure 1-6e) using microscopy imaging. Imaging techniques [10] commonly used by HT/HC screen are the following: (1) fluorescence microscopy [20][21][17][14] (cf. Figure 1-6a-c), (2) confocal laser scanning microscopy (CLSM) [22][23] (cf. Figure 1-6d), and (3) total internal reflection fluorescence (TIRF) microscopy [24][25]. When being applied for live cell imaging, the design is often referred as a time-lapse imaging (cf. Figure 1-4) for a temporal effect is studied. Similar to capturing video, time-lapse imaging captures images at a fixed sampling interval (temporal-resolution) from a predefined location in each well.

The temporal-resolution selected for time-lapse imaging, given in terms of the interval between two consecutive frames, is ranging from 30 minutes per frame for a stationary human reporting cell line [12] to 30 seconds per frame for matrix adhesion dynamics. The choice of the temporal-resolution is often considered empirical and subject-depended. The aim is to find a sample interval that captures the major changes of phenomenon. The temporal-resolution at which a phenomenon occurs usually does not require video rate. In that case, using higher temporal-resolution to capture additional intermediate stage does not necessarily provide more information. Moreover, sometimes to enforce a higher temporal-resolution may also produce undesirable quality issues such as cell death due to phototoxicity [26]. Therefore, the temporal-resolution of the imaging is tuned to take into account the preservation of cell vitality.

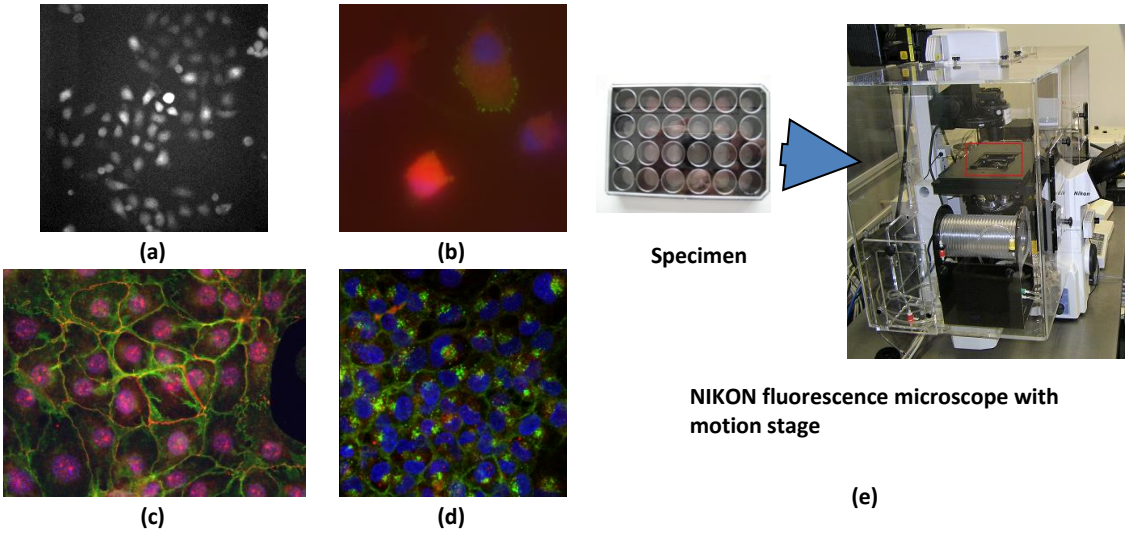

Figure 1-6 different image modalities and experiment designs in image acquisition (a) epifluorescence microscopy for cell migration analysis, (b) TIRF + epifluorescence microscopy for matrix adhesion analysis, (c) epifluorescence microscopy for protein localization during wound-&-recover, (d) confocal laser microscopy for protein translocation during cell endocytosis, (e) a simple setup of hardware for image acquisition

4. Image Analysis

The image analysis step (cf. **Chapter 2** and **Chapter 3**) is the crucial step in converting HT/HC images into numerical phenotypical descriptions. It is the bridge connecting the biological experiment and the data analysis. A robust image analysis solution is absolutely necessary to produce an objective understanding of the raw image data. In HT/HC screens, one has to deal with large quantities of images of which the quality may vary due to the fact that observations are done over a period of time as well as due to the variation in response of the cells. Consequentially, errors in image analysis will propagate into the subsequent steps. Therefore, it is important to use dedicated and problem-driven image analysis solutions. In general, our image analysis solutions for HT/HC screen studies consists of four major steps (cf. Figure 1-7): (1) **image enhancement** [27][28], (2) **image segmentation** [29][30], (3) **object tracking** [9][31], and (4) **phenotypical measurements** [10][32][33][14].

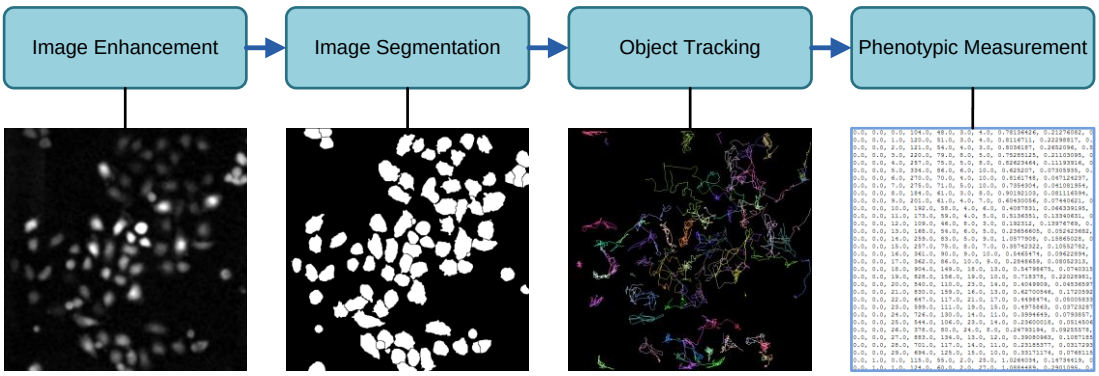

Figure 1-7 image analysis in HT/HC screen study

In the image analysis of HT/HC screen studies, the **image enhancement** serves to improve the separation between the foreground and the background. The image enhancement usually consists of two parts: (1) noise suppression and (2) signal enhancement. The noise suppression is a procedure to increase the signal homogeneity within foreground and background [34][35]. The signal enhancement is a procedure to increase the signal heterogeneity between foreground and background [34][35]. Image enhancement is not required if foreground and background can be well separated (see § 2.3). Otherwise, image enhancement is always recommended before image segmentation.

After image enhancement, the image will be segmented. **Image segmentation** is the procedure to convert the raw image into foreground and background. The foreground is defined as the pixels containing information for the analysis. In cytomics, image segmentation (cf. Ch. 2) is the essential step to convert the raw image into quantifiable elements. A good segmentation solution should correctly extract most of the foreground without introduction of background. Moreover, for HT/HC screen studies image segmentation algorithms are required to be self-adaptive and robust to image quality since it is impractical for users to tune the parameters for individual images. This is what generic segmentation algorithms often fail to comply. The image segmentation in HT/HC screen study will be further discussed in Chapter 2.

For a time-lapse imaging setting (cf. Figure 1-4), the objects from the segmented images will be tracked. The **object tracking** (cf. Ch. 3) is another crucial procedure in the image analysis to produce dynamic information of objects in a temporal study. In general, a tracking algorithm builds linkages between objects from consecutive frames using either motion models or object similarity measurements. In HT/HC screens, the efficiency of object tracking is often affected by the accuracy of image segmentation and the sampling interval of time-lapse imaging [9] [18][36]. The object tracking in HT/HC screen study will be further discussed in Chapter 3.

The phenotypic measurement is the final step of image analysis in an HT/HC screen study. It is accomplished by extracting numerical descriptors [9] using the object labeling from the image segmentation and object linkages from the object tracking. These numeric descriptors or measurements are information carried within the image data that can be used to characterize a biological phenomenon. With these measurements, data analysis can be performed to verify the initial research question.

5. Data Analysis

The data analysis step (cf. **Chapter 4** and **Chapter 5**) further augments the phenotypical measurements into comprehensive data representation and visualization (cf. Figure 1-7) by employing machine learning and statistical analysis [10][37][5][9][38][39]. The selection of data analysis solutions is depending on the experiment design. Commonly employed measures include first-order statistics, temporal low-order statistics, significance testing, data clustering and data classification. The data analysis theory is relative mature in other –omics studies, i.e. transcriptomics. In cytomics the analysis workflow is under development [29]. The decomposition and comparison of temporal data is not yet fully understood. The data analysis step is another important step, similar to image analysis, in converting image data into comprehensive conclusions.

1.3. Thesis Scope and Structure

This thesis is dedicated to the study of image analysis in HT/HC screens. A HT/HC screen produces an extensive amount of images for which manual analysis is impractical. Therefore, an automated image analysis solution should precede an objective understanding of the raw image data. The efficiency of HT/HC image analysis is often depending on image quality. Therefore, this thesis will address two major procedures in the image analysis step of HT/HC screens, namely image segmentation and object tracking. Moreover, in this thesis we will focus on extending computer science and machine learning approaches into the design of more dedicated algorithms for HT/HC image analysis. Additionally, this thesis demonstrates a practical implementation of an image and data analysis workflow case studies in which different imaging modalities and experimental designs are used. Finally, the thesis will briefly address an infrastructure for end-user interaction and data visualization.

This thesis is divided into three main parts:

Methodology

The first part includes "**Chapter 2 Robust Image Segmentation for Cytomics**" and "**Chapter 3 Robust Object Tracking for Cytomics**". It starts from a literature study of existing image segmentation and object tracking algorithms that have been frequently applied in HT/HC studies. Subsequently, it further elaborates the design and implementation of a dedicated image segmentation algorithm, namely watershed masked clustering, and two robust real-time object tracking algorithms for HT/HC studies, namely kernel density estimation with mean shift tracking and energy driven linear model tracking. In this part, the performance and efficiency of these algorithms are assessed using ground truth image sets from empirical live cell HT/HC screens. Moreover, the performance assessments of these algorithms are used to the case study part as a theoretical foundation of designing image analysis workflows for experiments.

Case Studies

The second part including "**Chapter 4 A Study to Cell Migration Analysis**" and "**Chapter 5 A Study to Dynamics of Matrix Adhesion**", demonstrates two case studies of different experiment designs and image modalities. The first case study is an analysis aiming to extract single-cell level phenotypical characterization from an aggressive cancer cell line with live cell

imaging. The second case study is matrix adhesion dynamics study. This project aims to model the correlation between the matrix adhesions, a type of subcellular macromolecular complex, and cell migration. This case study consists of a complex experiment setting and multi-modal live cell imaging.

These two case studies are demonstrating the practical implementation of image analysis and data analysis workflows following the design studies described in Chapter 2 and Chapter 3.

Supplementary System

The third part includes "**Chapter 6 HT/HC Data Management System**", illustrates a data management system to support HT/HC image analysis. Through the analysis pipeline, researchers often focus more on the image analysis and data analysis whilst an equal attention should also be given to the design of an efficient method to organize and visualize data. Moreover, HT/HC is but just one step in the cytomics pipeline. It is more important to integrate different -omics analysis to provide comprehensive understandings of biological phenomena. Thus, the design and implementation of a data management system offers a mutual platform for the data exchange between different –omics studies.