HIGH CONTENT ANALYSIS WITH CELLULAR AND TISSUE SYSTEMS BIOLOGY: A BRIDGE BETWEEN CANCER CELL BIOLOGY AND TISSUE-BASED DIAGNOSTICS

Albert Gough¹,², Timothy Lezon¹,², James R. Faeder¹, Chakra Chennubhotla¹,², Robert F. Murphy³, Rebecca Critchley-Thorne⁵ and D. Lansing Taylor¹,²,⁴

¹Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260

²University of Pittsburgh Drug Discovery Institute, University of Pittsburgh, Pittsburgh, PA, 15260

³Departments of Biological Sciences, Biomedical Engineering and Machine Learning and the Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, 15213

⁴University of Pittsburgh Cancer Institute, Pittsburgh PA 15213

⁵Cernostics, Inc. Pittsburgh, PA
Outline:

1) High Content Analysis (HCA)
   1.1) Background on fluorescence imaging in cancer biology, drug discovery and diagnostics
   1.2) Imaging live cells and model organisms with HCA
   1.3) Imaging fixed cells and model organisms with HCA
   1.4) Multiplexed to hyperplexed fluorescence-based HCA

2) Cellular Systems Biology in cancer research and drug discovery
   2.1) Cell lines, primary cells and tumor micro-environments in HCA studies
   2.2) Heterogeneity as a challenge in cancer biology and drug discovery

3) Tissue systems biology in cancer diagnostics/prognostics
   3.1) From H&E staining to multiplexed and hyperplexed fluorescence
   3.2) Heterogeneity and complexity of tumors: Breast cancer as an example
   3.3) Integration of multiple molecular data sets
   3.4) Tools to address heterogeneity in tissue systems

4) Computational and systems biology methods in cancer research and diagnostics
   4.1) Machine learning and other computational methods in cancer research, drug discovery and diagnostics
   4.2) Computational modeling of cellular and tissue systems
Abstract

Tumors are integrated systems of interacting malignant cells, stem cells and stromal components, including immune cells, fibroblasts, endothelial cells and nerve cells that evolve within the human tissues. Cancer displays significant genetic and non-genetic heterogeneity in time and space within tumors, as well as between patients. The complexity of the temporal and spatial aspects of cancer progression for diagnostics and therapeutics is daunting. However, the implementation of High Content Analysis (HCA) with the tools of computational and systems biology to investigate and to explain the complexities from 2D and 3D models of disease, to the stratification of patient tumor samples, is beginning to shed some light on the molecular basis of cancer. In this review, we will explore the integration of advances in applying HCA, computational biology and systems biology to the goal of creating better approaches to finding personalized therapeutics with companion diagnostics.
1) **High Content Analysis (HCA)**

1.1) **Background on fluorescence imaging in cancer biology, drug discovery and diagnostics**

Cancer displays significant genetic and non-genetic heterogeneity.\(^1\)\(^-\)\(^4\) Tumors are integrated tissue systems of interacting malignant cells, stem cells and stromal components, including immune cells, fibroblasts, endothelial cells and nerve cells (Figure 1).\(^5\)\(^-\)\(^9\) Immune cells and other stromal components play cooperative roles in tumor development and metastasis and influence responses to therapies. The stromal phenotype and functions are strongly associated with disease progression and clinical outcome in cancer. Leukocytes are attracted into tumors by chemokines and can both protect the tumor from anti-tumor immunity and promote tumor progression via stimulating angiogenesis and tumor cell migration.\(^10\)\(^-\)\(^12\) Tumors can render infiltrating immune cells anergic/non-responsive\(^13\)\(^,\)\(^14\) or drive such cells into apoptosis.\(^15\) Tumor cell heterogeneity, the complexity of the tumor system and the vital interactions of tumor cells with multiple components of the stroma highlight the need for a “tissue systems biology” approach to cancer diagnostics, which combines multiplexed biomarker measurements in the context of the tissue architecture and tumor cell function (see Table 1) with informatics tools to classify individual patients according disease subtype, recurrence and responses to therapies.\(^16\)

![Figure 1. Solid tumors are systems within the human system.](image) The tumor system includes normal tissue/organ cells (pink), tumor cells usually containing a range of genetic alterations (yellow), cancer stem cells (orange), along with immune cells (e.g. dendritic cells, granulocytes, macrophages, lymphocytes) and other stromal cells, such as fibroblasts and vasculature, that all play a part in tumor development, progression, metastasis and response to therapy, based in part on their spatial relationships.
The use of fluorescence-based imaging technologies have been applied to models of cancer and patient samples for many years. The applications have spanned the range from *In Vitro* studies in single cells, populations of cells, mixed cell populations, 3D tumor models, and pathology of patient tumor samples, as well as imaging cells within mouse tumor models, including anti-tumor immune responses, dynamics of cancer growth and invasion, tumor angiogenesis and regression, and tumor cell movements. Furthermore, applications in drug discovery have been performed in cells and in small experimental organisms including yeast, C. elegans, Drosophila and zebrafish, as well as monitoring tumors in rodent models with whole body imaging of small mammals.

The present chapter will focus on the application of HCA to populations of cells, more complex tumor models, *In Vitro* and patient samples where large image datasets can be created and
explored with computational and systems biology tools to create a bridge between cancer cell biology and tissue-based diagnostics. The investigation and integration of the continuum of single cells, cell populations, 3D tumor models and patient samples is needed to define the molecular basis of cancer.

High content analysis (HCA), originally termed High Content Screening (HCS), is a platform technology created in the 1990's to automatically image, analyze, store and mine large image datasets based primarily on fluorescence imaging microscopy, although transmitted light is an option.\textsuperscript{61,62} HCA harnesses advances in automation of microscopy, image processing, image analysis, fluorescence-based reagents, automation of sample preparation and relational databases (Figure 2).\textsuperscript{61-64} There have been numerous books\textsuperscript{65-67} and reviews on the applications of HCA in basic biomedical research, drug discovery/development and diagnostics.\textsuperscript{39,68-72} A broad range of fluorescence based reagents for both live cell, kinetic studies and fixed end-point investigations have also been reviewed in detail.\textsuperscript{63} The major types of reagents, readouts and selected on-line databases are listed in Table 2. Extensive lists of additional reagent sources can be found online\textsuperscript{73,74} and in published catalogs.\textsuperscript{75}

\textbf{Figure 2. The Components of High Content Analysis (HCA).} HCA is defined by the integration of: A) Arrays of cells/tissues for high throughput biology; B) Automated microscope systems available from multiple vendors; C) A wide range of reagents and cell types; D) Automated sample preparation systems and protocols; E) Imaging algorithms typically designed to measure multiple features for each cell; F) Informatics to review and further process the data, for example to fit dose response curves; G) Bioinformatics to relate multiparameter cellular features to biological functions.
Table 2: Classes of fluorescence-based reagents, readouts and on-line resources for HCA.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Type</th>
<th>Application</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluorescent Probe Classes</td>
<td>Chemical fluorophores</td>
<td>Wide spectral range, easily attached to targeting molecules, some are environmentally sensitive, many useful properties</td>
<td>61,63,74,75</td>
</tr>
<tr>
<td></td>
<td>Nanocrystals</td>
<td>Stable, bright, single excitation, narrow emission, best for multiplexing in fixed cells or cell surface markers</td>
<td>75,110,332</td>
</tr>
<tr>
<td></td>
<td>Fluorescent Proteins (FPs)</td>
<td>Multiple wavelengths, transient or stable expression, linked to targets, some are environmentally sensitive, photoactivated (or switched), live or fixed cell assays</td>
<td>63,69</td>
</tr>
<tr>
<td>Antibodies</td>
<td></td>
<td>Target expression level and localization</td>
<td>63,69,75</td>
</tr>
<tr>
<td>FISH Probes</td>
<td></td>
<td>DNA copy number variants, RNA expression, including micro-RNA</td>
<td>197</td>
</tr>
<tr>
<td>Fluorescent Proteins (FPs)</td>
<td></td>
<td>Target expression level, localization and dynamics, photobleaching or photoactivation for transport within or between compartments</td>
<td>63,69</td>
</tr>
<tr>
<td>Environment Sensitive Probes</td>
<td></td>
<td>Ion concentrations, membrane potential, hydrophobic compartments</td>
<td>63,69,75</td>
</tr>
<tr>
<td>Proximity Probes</td>
<td></td>
<td>FRET, co-localization</td>
<td>63,69,75,333,334</td>
</tr>
<tr>
<td>Enzyme Activity</td>
<td></td>
<td>Fluorogenic substrates, cleavable linkers</td>
<td>63,75</td>
</tr>
<tr>
<td>Organelle Specific</td>
<td></td>
<td>Nucleus (DNA), mitochondria, lysosomes, neutral fat, endoplasmic reticulum, etc.</td>
<td>63,75</td>
</tr>
<tr>
<td>Fluorescent Biosensors</td>
<td></td>
<td>FPs or combinations of FPs engineered to report on activation of biomarkers or pathways</td>
<td>63,69</td>
</tr>
<tr>
<td>HCA Readouts</td>
<td>Intensity</td>
<td>Relative concentration of target</td>
<td>63,69</td>
</tr>
<tr>
<td></td>
<td>Distribution</td>
<td>Distribution and dynamics of molecular targets in cells</td>
<td>63,69</td>
</tr>
<tr>
<td></td>
<td>Co-localization</td>
<td>similarity or difference in the distribution of 2 or more labels</td>
<td>63,69</td>
</tr>
<tr>
<td></td>
<td>FRET</td>
<td>Very sensitive determination of close proximity of 2 labels</td>
<td>63,69</td>
</tr>
<tr>
<td></td>
<td>Morphology</td>
<td>Texture, size or shape of cells or organelles, aggregation</td>
<td>16,68</td>
</tr>
<tr>
<td></td>
<td>lifetime</td>
<td>Local chemical environment</td>
<td>63,69</td>
</tr>
</tbody>
</table>
Although the development of HCA has focused on the application of fluorescent probes, chromogenic probes continue to be used extensively for labeling tissue sections. Table 3 compares the advantages and disadvantages of fluorescent and chromogenic probes. Most HCA systems are optimized for the use of fluorescent probes, principally for their high sensitivity, high specificity, broad range of cellular functional readouts, broad range of wavelengths for multiplexing, and the ability to engineer cells to express fluorescent proteins and biosensors. Since HCA makes use of automated imaging and quantitative image analysis, there is no need for directly viewing the labeled specimen, and once the images are acquired; there is no further need for the specimen other than for institutional or clinical requirements. In traditional pathology on the other hand, chromogenic probes have some advantages. The human brain is still the most sophisticated and reliable image processor for interpretation of small numbers of images. Readily available, low cost chromogenic probes provide a stable and dense labeling for visualization in a transmitted light microscope or by digital image pathology, while simultaneously viewing the contextual morphology of the cells. Although providing somewhat lower resolution and more limited multiplexing than fluorescent probes, chromogenic probes still provide a good labeling strategy where one to three biomarkers per slide can be useful.
Table 3. Comparison of Fluorescent and Chromogenic Readouts 177,341-344

<table>
<thead>
<tr>
<th>Reporter Type</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluorescent</td>
<td>Present standard in cell analysis</td>
<td>Reagents are less stable for long term storage</td>
</tr>
<tr>
<td></td>
<td>High sensitivity and specificity</td>
<td>More expensive fluorescence more expensive imaging systems</td>
</tr>
<tr>
<td></td>
<td>Quantitative readout</td>
<td>More expensive reagents</td>
</tr>
<tr>
<td></td>
<td>Multiplex targets that are co-localized and/or in close proximity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Broad spectrum of wavelengths</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Higher resolution with confocal imaging</td>
<td></td>
</tr>
<tr>
<td>Chromogenic</td>
<td>Present standard in tissue analysis</td>
<td>Variable sensitivity and specificity</td>
</tr>
<tr>
<td></td>
<td>Long term stability of labeling</td>
<td>Multiplexed targets must be spatially separated</td>
</tr>
<tr>
<td></td>
<td>Brightfield Microscopes</td>
<td>Precipitates cause fuzziness around target</td>
</tr>
<tr>
<td></td>
<td>Greater amplification</td>
<td></td>
</tr>
</tbody>
</table>

Success in the human genome project demanded tools to define the functions of the coding and non-coding portions of the genome, to define the dynamic interplay of cellular constituents within and between cells, to characterize sub-populations, as well as to define the relationships between populations of cells in higher order biological systems. The field was named "cellomics" and the platform technology was named HCA. HCA harnesses the ability to implement combinatorial treatments on large sample sizes by using microplates, patterned microarrays 77 and microfluidic devices 78 for cells, microplates for small organisms and mounted sections/microarrays for tissues. These large sample sizes are required for statistical analyses and the exploration by computational and systems biology. 16,21,63,79,80
1.2) Imaging live cells and model organisms with HCA

Imaging living cells and model organisms by HCA has the advantage of allowing the investigation of the dynamic, temporal-spatial interplay of cellular constituents that define normal and abnormal cell and tissue functions. Either single time points and/or kinetic measurements can be generated and analyzed. It is also possible to harness advanced, fluorescence-based probes and biosensors to measure physiological parameters not readily measured in fixed samples, such as cyclic protein translocations, pH, free Ca++, membrane potentials and a growing number of physiological biosensors by fluorescence microscopy. A disadvantage of investigating living systems by HCA is that biological processes can change from the time of imaging the first well in a microplate to the last well and this issue is multiplied when going from 96 to higher density well plates. Depending on the time course for the specific biological process, including cyclic changes and the protocol for the addition of experimental treatments, large-scale, living samples usually have to be profiled in smaller batch sizes.

All HCA profiling or screening studies start with living systems that receive some combinatorial application of small molecules or biologics, RNAi for knockdowns and/or nucleic acids for transfections or transductions. Although more demanding to perform, a recent investigation studied the kinetics of response of individual cells to drug treatments demonstrating the variability of cellular responses in a population. Measuring kinetic responses should increase as even more biosensors are developed and the complex and dynamic aspects of signaling processes are investigated.

1.3) Imaging fixed cells and model organisms with HCA
The main advantage of using fixed samples is that large-scale sample preparation and robotic screening of many microplates or slides is possible without changes in the biology during the read-out. Therefore, many combinatorial treatments can be prepared at one time and the microplates/slides stacked in a robotic system for screening/profiling. There are many fluorescence-based reagents including antibodies, fluorescence in situ hybridization (FISH) probes and fluorescent proteins that can be used to define single time point localizations, relative concentrations and activities. In order to optimally interpret fixed samples, the half-time of a process under investigation must either be determined in live sample profiles or multiple time points must be generated in distinct wells or plates.

HCA has been extensively applied as a phenotypic approach to cancer drug discovery over the last few years, in both primary and secondary screens, either using live cell or fixed cell screening. Although specific molecular targets guide many of these screens, pathway modulations and phenotypic profiling is central to the approach. Examples of cancer biologies explored include energy metabolism, viral induction, apoptosis, cell cycle, autophagy, tumor invasion and metastasis, pathway modulations, a panel of biologies and phenotypic changes compared to mutants. In many cases, HCA is also used in structure activity relationship (SAR) to optimize lead compounds. However, it is still important to ultimately identify the mechanism(s) of action of lead compounds. The role of HCA in cancer drug discovery and development has been further advanced with the application of more quantitative analyses of profiles using computational biology and systems biology approaches, and will be explored in detail below.
1.4) **Multiplexed to hyperplexed fluorescence-based HCA**

It has been the goal of imaging cytometry to increase the number of specific molecular parameters that can be measured in the same sample, so that complex interplays of components, pathway mapping and heterogeneity of biological processes can be analyzed in increasing detail. We have defined multiplexed fluorescence in imaging applications as the combination in a sample of up to seven fluorescent probes that can be discriminated by spectral selection. Multiplexing has been accomplished in both live and fixed samples using a range of fluorescent probes. Multiplexing by flow cytometry has reached the level of 15-18 distinct fluorescent probes, but flow cytometry does not permit analyses of the temporal-spatial dynamics within or between cells. Hence, imaging technologies are being advanced to produce more parameters, especially in fixed samples. There have been a number of technical developments to increase the number of fluorescently labeled antibodies and FISH probes per sample, including new types of probes like quantum dots, new algorithms like spectral unmixing, and new protocols like sequentially labeling, imaging, quenching the fluorescence, and repeating the process. Recently, the GE Global Research Center has demonstrated that greater than 60 fluorescence-based biomarkers can be applied to a single tissue sample using a sequential labeling approach. This novel platform technology should greatly impact basic cancer research, drug discovery and diagnostics/prognostics. Generating the multiplexed to hyperplexed data sets creates a powerful platform that will enable the application of advanced computational methods to directly define pathways and modifications due to perturbations, as well as to characterize and to understand heterogeneity. It is also possible to harness fluorescence lifetime imaging to gain some parameters, as well as the application of mass spectroscopy applied to single cells and tissues, but these latter approaches are not covered here.
In addition, other imaging modalities have been applied to cancer model systems and patients.\textsuperscript{33,119,120} The data from these investigations must be integrated into the systems biology models developed in cells, small experimental organisms and patient tissue sections.

2) Cellular systems biology in cancer research and drug discovery

2.1) Cell lines, primary cells and tumor micro-environments in HCA studies

The biological interpretation of cancer genomic and proteomic data remains a major challenge.\textsuperscript{121} The presence of both genetic and non-genetic heterogeneity among cancer cells complicates population level data interpretation and implies the need for more detailed analysis at the cellular level.\textsuperscript{122-125} Furthermore, the development of diagnostics and therapies requires consideration of the functions and responses in all cancer cells, not just the ‘average’ cancer cell, as well as the role of the microenvironment. There has now been a shift in focus on screening for cancer drugs from relying exclusively on tumor cytotoxicity to understanding the signaling context within which the particular molecular target operates.\textsuperscript{126} Multiplexed or hyperplexed cell analysis, when combined with computational modeling, can serve to directly assess multiple functions at the cellular level. Cell-by-cell analysis also allows detailed determinations of heterogeneity in cell populations. Rather than viewing heterogeneity as an interference to the development of diagnostics and therapies, computational models of heterogeneity can provide a deeper understanding of the functioning of the underlying pathways and networks, as discussed below.\textsuperscript{127} Furthermore, HCA can accelerate mechanism of action studies for oncology research,\textsuperscript{128} provided relevant models are used in the analysis. Two important aspects of cancer model development are the choice of cells and the design of the model.
There are now a large number of cancer cell lines available; however, the relevance of those cell lines has long been debated.\textsuperscript{129-132} Cancer cell lines are attractive models as they provide an unlimited source of homogeneous, self-replicating material, free of contaminating stromal cells, and the majority are easy to culture in standard media. Although cancer cell lines may have diverged to some extent during culture, the genomics and even pharmacological profiles of large numbers of these cell lines are available, and provide a basis for interpreting cellomics profiles in the context of pathways derived from genomics. For example, the recently published Cancer Cell Line Encyclopedia (CCLE), a compilation of gene expression, chromosomal copy number and massively parallel sequencing data from 947 human cancer cell lines, coupled with pharmacological profiles for 24 anticancer drugs across 479 of the cell lines\textsuperscript{133} extends existing data characterizing cancer cell lines.\textsuperscript{134-137} Furthermore, results from studying cancer cell lines in 2D serve as the simplest model for comparison with results from more complex 3D models involving tumor microenvironments, as well as results from patient samples, such as those available from The Cancer Genome Atlas.\textsuperscript{138}

Cell line fidelity with primary tumors varies depending on the tumor type. For example, the fidelity of breast cancer cell lines with primary tumors, assessed by comparison of genetic heterogeneity and copy number abnormalities, has been shown to be high\textsuperscript{134}. Conversely, cell lines derived from glioblastomas have lower fidelity with primary tumors and mostly lack amplification and mutation of EGFR that is found in approximately 50% of glioblastomas.\textsuperscript{139} While it is clear that cell lines cannot model the hundreds to thousands of genetic aberrations found in primary tumors, HCA enables sufficient assessment of the many potential aberrations to identify those that are functional.
As a group, breast cancer cell lines faithfully reproduce the heterogeneity in human breast
tumors as described above, though individually they exhibit profiles that fall short of truly
representing the intratumoral heterogeneity of individual breast tumors. Studies suggest that
collections of cell lines representing multiple cell types can be used to model the cellular
heterogeneity in tissues. Individual cell lines can also be useful models for specific functions.
For example, the most commonly used breast cancer cell line, MCF-7 established in 1973 at the
Michigan Cancer Foundation, exhibits exquisite hormone sensitivity through expression of
estrogen receptor, making it an ideal model to study the hormone response. It is important then
to understand how well and which cell lines best model the diversity and which best model
specific pathways. Microarray studies have identified molecular subtypes—luminal A, luminal B,
ERBB2-associated, basal-like and normal-like—with characteristic gene expression patterns and
underlying DNA copy number alterations (CNAs). Genomic profiling of a collection of breast
cancer cell lines found that they retained expression patterns with relevance to the luminal-basal
subtype distinctions. That compendium of molecular profiles defines cell lines suitable for
investigations of subtype-specific pathobiology, cancer stem cell biology, biomarkers and
therapies, and provides a resource for discovery of new breast cancer genes. The choice of
cancer cell lines, primary cells or stem cells to construct models for high content analysis should
be made based on the basis of the abundant genomics, proteomics and pharmacological profiles
of the cell model, the goals of the study and the analysis methods to be employed.

While 2D cultures of cell lines provide simple models for HCA, pathway and network modeling
(see below), and drug discovery, more sophisticated 3D microenvironment models are
required to better recapitulate the tumor environment. 3D cell cultures are rapidly becoming the
method of choice for the physiologically relevant modeling of many aspects of non-malignant
and malignant cell behavior \textit{ex vivo}. 3D models include: relatively simple tumor spheroid models;\textsuperscript{144,145} more complex extracellular matrix models;\textsuperscript{146-149} models which mimic the architecture of specific cancers;\textsuperscript{150} cells grown on artificial scaffolds such as engineered organ models;\textsuperscript{151-154} and tissue explants.\textsuperscript{155,156} In addition, development of microfluidic devices capable of maintaining a controlled, physiological environment will allow the construction of even more relevant tissue microenvironment models.\textsuperscript{157}

In general, the simpler 2D models allow the greatest flexibility in terms of the range of markers and measurements that can be made cell-by-cell in the monolayer, using multiplexed HCA. Although confocal HCA systems can optically slice through thick specimens, the fact that the cells are not all in the same plane complicates the cell-by-cell quantitation of features. The approaches to the analysis of confocal sections are similar to the analysis of tissue sections, which is discussed in more detail below. The spheroid model is arguably the simplest 3D model for studying tumor cell biology, therapy resistance, cell-cell interactions, invasion, drug penetration, modeling, nutrient gradients, and tumor cell metabolism.\textsuperscript{144,145} The self-organization and generation of distinct tumor microenvironments makes it an attractive model for high throughput imaging.\textsuperscript{158} The well-defined geometry of the spheroid simplifies the identification of microenvironments in the model.

There are a wide range of 3D tissue models that can be used with HCA analysis. Extracellular matrix models of various compositions attempt to reconstruct the \textit{in vivo} environment for studies of tumor cell biology, cell-cell interactions, cell migration and invasion. Layers of cells are cultured on top of porous membranes for drug transport and binding, therapy resistance, invasion assays. Engineered models such as micropatterned surfaces and 3D organizations,\textsuperscript{147,150} as well as cells cultured within a network of perfused artificial capillaries for studying tumor cell
metabolism, therapy resistance and artificial organs are the future. Each of these models, combined with a carefully chosen cell type and analysis approach, can provide insights into many functional components of cancer.\textsuperscript{159} A deeper systems biology understanding of cancer will come from integrating data across these studies through the construction of computational models, based on combining pathways based on genomics, with functional data from HCA studies, as discussed below.

2.2) Heterogeneity as a challenge in cancer biology and drug discovery

Cell-to-cell differences are always present to some degree in any cell population, and therefore the ensemble behavior of a population may not represent the behaviors of any individual cell.\textsuperscript{160} Broadly speaking, cellular heterogeneity can be classified as either genetic or non-genetic in nature.\textsuperscript{125} Genetic heterogeneity has been identified as an important factor in cancer progression, and is thought by some to result from pressure for cells to adapt to new environments during metastasis or to evade immune responses.\textsuperscript{161} Single-cell PCR gene-expression analysis, genomic profiling, immunohistochemistry and other methods show that cancer tissues contain distinct cell populations and that the different gene expression programs linked to multilineage differentiation are strongly associated with patient survival.\textsuperscript{162} Although genetic heterogeneity across cancer lines can be used as a predictor of drug efficacy,\textsuperscript{137} intratumor genetic heterogeneity can also impede development of personalized medicine. Phylogenic differences, both intratumor and between primary tumors and their metastases, indicate that the sequence of a single sample does not reveal the full complexity of tumor genetics.\textsuperscript{2}

Non-genetic heterogeneity is thought to originate from natural biochemical variations between clonal cells, such as differences in concentrations of biological compounds or slight discrepancies in the timing of cellular events. It plays an important role in cancer progression, as
subpopulations of cells with significant stable variations in biochemistry exist within primary and metastatic tumor cell populations. Transient drug-resistant phenotypes have been observed to emerge in cancer, conferring drug resistance in the absence of genetic mutation, and drug treatment has even been observed to induce novel phenotypes. The non-genetic heterogeneity in cancer further complicates therapeutic development, as it implies that genetic information by itself may be insufficient to explain the response of a particular tumor to treatment, even if distal regions of the tumor are sequenced. An emerging course of action is to address cancer at the level of pathway, simultaneously targeting multiple pathways to minimize potential drug resistance. In addition, defining the immune status and stromal cell content in tumors is critical to understanding individual responses to therapeutics.

Analyzing single cells within a population is a new frontier in platform technologies that has the potential to transform systems biology through new discoveries derived from cellular heterogeneity. A great amount of information about the biochemical and environmental conditions of genetically homogeneous cellular populations can be uncovered by exploiting the differences between individual cells within the population. Patterns of heterogeneity in basal populations have been used to predict drug sensitivity in clonal populations of the NCI-60 panel of cancer cell lines, and phenotype profiling has been applied to uncover compound mechanisms of action.

The general approach to extracting biochemical information from inherent heterogeneity commences with identification of well-defined phenotypes, either through expert opinion or automated clustering algorithms. In the case of the former, supervised machine learning techniques are used to teach a computer how to correctly recognize phenotypes and classify cells. It has been shown that a computer trained by an experienced biologist to recognize 14
distinct morphological phenotypes can accurately classify human cancer cells at a rate of 70,000 per second. Alternatively, unsupervised learning methods can be employed to allow computational identification of phenotypes based on clustering in feature space, without prior human interpretation. As an example, cellular phenotypes that were automatically identified using a support vector machine have been used to extract information about drug activity in HeLa cells. In either case, it is necessary to use a set of features that is capable of distinguishing between phenotypes. Once phenotypes are established and recognizable by computer, entire cell populations can be described in terms of the relative abundances of their constituent phenotypes. Statistical or mechanistic techniques are used for analysis of the patterns of heterogeneity. Statistical analysis techniques involve comparing phenotype distributions that result from novel treatments to those that are associated with treatments of known action. Standard statistical measures, such as the KS statistic or KL divergence, are used to compare population distributions, and predictions for novel treatments are based on which known treatments they most closely resemble. In mechanistic analysis, one or more cellular pathways of interest are modeled in silico, and their predictions are compared to the patterns of heterogeneity in the population. Predictions of concentrations of biomolecules within cells from pathway models can be compared directly to measurements from HCA experiments, and multiple pathway models can be combined to computationally represent a heterogeneous population of cells.

Figure 3 (left) illustrates the automated clustering of Cal33 head and neck cancer cells based on the relative activation of STAT1 and STAT3 in response to treatment with IFNγ and IL6 as measured by phosphorylation or translocation to the nucleus. In many cancers, abnormal activation of STAT3 functions to promote tumor growth, while STAT1, despite a greater than
50% sequence homology, is mostly antagonistic to STAT3 activity. A better understanding of the differences in regulation and activation of STAT1 vs. STAT3, on a cell-by-cell basis, will be important for therapeutic development. The cells were clustered on the STAT1 and STAT3 activation state using a Gaussian mixture model. The fraction of cells in each cluster varies over time, with a rapid but transient activation of STAT3 followed by a slower, but more sustained activation of STAT1 (Figure 3, right). The extent of each cluster indicates a highly heterogeneous level of activation, and the inclusion of a fraction near the origin indicates cells that appear to be unresponsive. Whether this is a kinetic effect or the result of the cells being in an activation resistant state is still under investigation.

Figure 3. Automated Identification and Quantitation of Subpopulations of Cells. A) The scatter plot illustrates the relationship between STAT3 and STAT1 activity for ~72,000 Cal33 head and neck cancer cells stimulated with 50ng/mL of IL6 and 50ng/mL of IFNγ for 8-120min. STAT3 activity measured as fluorescence intensity of an antibody to phospho-STAT3-Y705, and STAT1 activity measured as the nuclear-cytoplasmic difference in fluorescence of an antibody to STAT1. Five subpopulations (blue, green, yellow, orange and red ellipses) were automatically identified by a Gaussian mixture model algorithm. Pseudocolor images (inset), STAT1 (green channel), STAT3 (red channel) show the variation in labeling seen between the 5 clusters. B) The distribution of the cells between subpopulations changes over time (colored lines correspond to clusters). At the earliest time point, 8 min, over 60% of the cells are in cluster 5 (blue), principally STAT1(-) and STAT3(+). At 15 min, the largest fraction has shifted to cluster 4 (green) showing a slight activation of STAT1 and already a decrease in the STAT3 maximum. At 30min the cells are about equally distributed between clusters 3 (yellow), 2 (orange) and 1 (red). By 1 hour, clusters 1 and 2 account for >95% of all the cells.
3) Tissue Systems Biology in Cancer Diagnostics/Prognosis

3.1) From H&E staining to multiplexed and hyperplexed fluorescence

For decades pathologists have used Hematoxylin and Eosin (H&E) and immunohistochemistry (IHC) with manual analysis by microscopy to assess tissue morphology and protein expression and distributions for diagnostic purposes. While these methods are very valuable and are still used in the majority of cancer diagnoses, they are limited by subjectivity, intra- and inter-observer variability, and are only semi-quantitative (Table 3). There is a growing trend to digitize pathology to improve objectivity, standardization and productivity, as has been done in radiology. Brightfield and fluorescence digital slide scanners from companies such as Aperio Technologies,175 BioImagene (now Ventana Medical Systems),176 Perkin Elmer, Omnyx and 3DHistech177 are increasingly being adopted in medical laboratories. Slide scanners produce whole slide digital images for viewing and sharing by telepathology and to enable more reproducible analysis of morphology and biomarkers by image analysis software. Digital image analysis has been applied to IHC-labeled biomarker analysis in multiple types of cancer to improve reproducibility and standardize scoring methods.178-181 Even with digital slide analysis, the accuracy of chromogenic IHC is limited by the inherent staining variability associated with the non-linear signal amplification and is mostly limited to a few biomarkers per slide (Table 3).182,183 Multiplexed IHC can be performed,184 yet multi-chromagen images are more difficult to interpret visually and to deconvolve with image analysis software, particularly for biomarkers expressed in the same subcellular compartment (Table 3). However, there is now some promise in applying unsupervised learning methods to chromagen labeled samples.
Quantitative, multiplexed digital imaging methods can improve reproducibility and also measure the multitude of molecular changes in cancer that are associated with disease subtype, tumor progression and response to therapies. Since these molecular changes precede morphological changes in the development and progression of cancer, molecular profiling can improve the sensitivity of diagnostic and prognostic testing in conjunction with histopathology. Cancer progression and responses to therapies depend on multiple components of many signaling pathways within malignant cells and stromal processes. It has become clear that there are no “single bullet” biomarkers and that multiple biomarkers are required to accurately diagnose, predict risk of recurrence and/or response to therapies for all patients with a particular type of cancer. Multiplexed fluorescence biomarker labeling with digital imaging represents a significant improvement over traditional histologic methods for tissue biomarker analysis, including the ability to quantify multiple antigens per tissue section, the ability to quantify antigens that are colocalized to the same subcellular compartment, and more consistent, linear, higher resolution labeling with greater dynamic range of biomarker measurements. Multiplexed biomarker analysis in the context of tissue morphology builds on cellular HCA, described above, and allows objective, reproducible extraction of quantitative biomarker data and morphology data by image analysis software. Quantitative multiplexed fluorescence analysis of biomarkers in digital tissue images has been used to measure diagnostic, prognostic and predictive cancer biomarkers in multiple cancer types. These methods of protein biomarker measurement correlate with Western blot analysis, match or exceed the accuracy of manual IHC scoring and can improve standardization of tissue biomarker measurements. Labeling and imaging of nucleic acids in tissues can also add power to cancer testing. Locus-
specific amplifications and losses and also microRNAs can be measured by FISH in intact tissues to enable interpretation within the context of tissue architecture.\textsuperscript{197}

Multispectral imaging enables separation of overlapping dyes in both brightfield and fluorescence tissue images to increase the potential level of multiplexing.\textsuperscript{177} Nanoparticles such as quantum dots and composite organic-inorganic nanoparticles (COINs) have emission spectra that are narrower and more symmetrical than traditional fluorophores, which results in minimal crosstalk between fluorescence channels and offers the potential for higher levels of multiplexing in tissues.\textsuperscript{110,198} Hyperplexed fluorescence methods, defined above, can overcome the spectral limitations of fluorescence-based detection of biomarkers and offer the potential to assess dozens of protein and nucleic acid parameters per tissue section for protein network topology, spatial mapping of protein clusters, microRNA’s and copy number variants (CNV) in tissues.\textsuperscript{114-116,199}

Infrared spectroscopic imaging and Raman scattering imaging of unlabeled tissue sections enables collection of thousands of spectra representing biochemistry in the context of the tissue architecture.\textsuperscript{200,201} This type of high content tissue imaging can be used to classify cancer types and distinguish cancer stages;\textsuperscript{202-204} however, the mechanisms underlying the classifications are difficult to interpret since the specific molecules responsible for the hyperspectral signatures are not known.

3.2) Heterogeneity and complexity of tumors: breast cancer as an example

Tumor heterogeneity and complexity highlight the need for a “tissue systems biology” approach to cancer diagnostics, as described in section 1 and as exemplified by breast cancer. The current standard diagnostic and predictive tests for breast cancer are IHC-based measurements of estrogen receptor (ER), progesterone receptor (PR) and HER2/neu.\textsuperscript{205} These biomarkers are
usually measured one at a time in serial sections with manual scoring by a pathologist. HER2/neu status is confirmed by FISH in a subset of patients who are candidates for HER2/neu-directed therapies. Multiplexed fluorescence immunohistochemistry has been applied to these standard breast cancer biomarkers to improve standardization.\textsuperscript{206,207} This limited set of biomarkers is insufficient to address breast cancer heterogeneity and cannot accurately stratify patients according to breast cancer subtype, predict risk for recurrence, or benefit from therapies. Gene expression profiling has identified molecular portraits of the main breast cancer subtypes and characterized the genetic and epigenetic abnormalities associated with each subtype.\textsuperscript{1,3} RT-PCR- and DNA microarray-based multi-gene tests have been developed for prognostic and predictive testing in breast cancer. Oncotype Dx (Genomic Health, Inc., CA, USA) and Mammaprint (Agendia, BV, Netherlands) are two such tests that have achieved clinical adoption.\textsuperscript{208-212} These approaches are valuable and have had a positive impact on patient care; however, the various multi-gene tests for cancer can be limited by bias introduced by inter-patient variations in percentages of malignant, immune and stromal cells. Furthermore, these methods require tissues to be digested and the tissue architecture and spatial information, which are important for accurate biomarker measurement and interpretation, are lost. Stromal signatures including immune response and angiogenesis-related genes have been shown to not only add prognostic information but to have improved prognostic significance over standard breast cancer biomarkers and whole-tumor signatures in breast cancer.\textsuperscript{213} The rationale for profiling biomarkers by digital imaging in intact tissue sections is based on extensive literature describing the importance of protein level, activation status, sub-cellular localization, tissue localization, spatial relationships and distributions for the accurate measurement and interpretation of cancer biomarkers. Biomarker function is often more relevant to clinical
variables than overall expression. For example, transcription factors such as nuclear factor-κB (NF-κB) and signal transducers and activators of transcription (STATs) are ubiquitously expressed but are frequently activated in breast cancer. Their activation is measured by translocation to the nucleus and/or by phosphorylation. Examples of cancer biomarkers exhibiting subcellular relocation based on pathway activity are provided in Table 4 and tissue images showing nuclear translocation of transcription factors are shown in Figure 4.

Table 4. Examples of Cancer Biomarkers Exhibiting Subcellular Relocation based on Pathway Activity

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Subcellular Relocation based on Pathway Activity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>NF-kappaB</td>
<td>Restricted to cytoplasm in resting state, translocates to the nucleus upon activation and induces transcription of genes involved in inflammation, proliferation, apoptosis in malignant cells and stromal cells.</td>
<td>345,346</td>
</tr>
<tr>
<td>Beta-catenin</td>
<td>Activation of Wnt signaling leads to nuclear translocation of beta-catenin, which drives epithelial-mesenchymal transition and metastasis.</td>
<td>347-349</td>
</tr>
<tr>
<td>STAT1, STAT3, STAT5</td>
<td>Activation of STATs requires phosphorylation and nuclear translocation, which are associated with cancer prognosis</td>
<td>215,350,351</td>
</tr>
<tr>
<td>p21</td>
<td>Nuclear localization is correlated with the inhibitory effect of p21 on cancer cell growth, whereas cytoplasmic localization is associated with protection from apoptosis</td>
<td>352-354</td>
</tr>
<tr>
<td>HIF1-alpha</td>
<td>Master regulator of oxygen homeostasis that is cytoplasmic under normoxic conditions, translocated to the nucleus in response to hypoxia. Associated with clinical outcome in multiple cancer types.</td>
<td>355,356</td>
</tr>
<tr>
<td>FOXO3a</td>
<td>Tumor suppressor that is active when localized to the nucleus and inactive when in the cytoplasm. Nuclear localization is associated with good prognosis and cytoplasmic localization is associated with poor prognosis in breast cancer.</td>
<td>357</td>
</tr>
</tbody>
</table>
Figure 4. Multiplexed Immunofluorescence Labeling and Digital Imaging of Functional Biomarkers in Tissues. Sections of esophageal tissue with intestinal metaplasia (panel A) and breast cancer tissue (panel B) were labeled with a primary antibody specific to NF-κB p65, fluorescently-labeled secondary antibody and Hoechst 33342 (nuclear stain). Slides were imaged using a fluorescence digital slide scanner (Aperio FL, Aperio Technologies, Inc., CA) at 20x magnification. Examples of cells exhibiting resting state NF-κB in the cytoplasm are highlighted with purple arrows and cells with activated nuclear NF-κB with blue arrows (panels A and B).

Spatial relationships and microenvironments within the tumor system are also important in breast cancer. A three part arrangement of an invasive breast cancer cell, a macrophage and an endothelial cell in breast cancer has been described as a microenvironment conducive to distant metastasis. The combined density and spatial distribution of mature dendritic cells and activated T cells has also been shown to have prognostic significance in breast cancer.

Macrophages infiltrate only the stroma in subsets of breast cancer patients, and infiltrate the tumor nests in other patients (Figure 5 G-H), which has been correlated with survival in breast cancer patients. This valuable spatial information is lost in molecular profiling approaches that digest tissues to extract DNA, RNA or proteins, but is preserved and can be quantified by digital imaging approaches. Digital imaging of intact tissues also enables measurement of nuclear morphometric features and the amount of DNA in tumor cells, which have diagnostic and prognostic significance in breast cancer. Multiplexed to hyperplexed biomarker imaging is also important for measurement of specific cell types and phenotypes such as cancer stem cells and immune cell subsets that require multiple biomarkers for accurate classification.
A major goal of hyperplexing will be to identify the optimal combination of protein, DNA and RNA biomarkers that can stratify the disease subpopulations. Subsequently, a multiplexed subset of the biomarkers will be used in the development of diagnostic/prognostic tests. Figure 5 illustrates multiplexed immunofluorescence imaging of serial sections of a breast cancer tissue microarray, scatter plot analysis of the relationships between single cell features derived from the images and shows examples of immune cell subsets infiltrating the cancer stroma and tumor cell nests.
were labeled with primary antibodies specific for Ki-67, pan-cytokeratin, CD68, COX-2, NF-κB p65, PD-1 and CD45RO, fluorescently-labeled secondary antibodies specific to each primary antibody and Hoechst 33342 (nuclear stain). Slides were imaged using a fluorescence digital slide scanner (Aperio FL, Aperio Technologies, Inc., CA) at 20x magnification. The digital images were analyzed using TissueCIPHER software (Cernostics, Inc., Pittsburgh, PA) to segment nuclei, cells and tumor cells nests as individual objects in which quantitative biomarker intensity measurements were made. A: Ki-67 (green nuclear signal both in stroma and tumor nests), pan-cytokeratin (red plasma membrane signal in tumor nests) and DNA (blue nuclear signal) in breast cancer tissue core; B: CD68 (green plasma membrane signal in stroma), NF-κB p65 (red cytoplasmic and nuclear signal throughout tissue), COX-2 (yellow plasma membrane and cytoplasmic signal throughout tissue) and DNA (blue nuclear signal); C: Nuclear image analysis mask; D: Tumor nest image analysis mask; E: Single cell measurements of cellular CD68 and COX-2 in the breast cancer tissue shown in image B. The bivariate data shows that the tissue is composed of COX-2-positive, CD68-negative malignant cells and CD68-positive cells, i.e. macrophages, with a subpopulation expressing high levels of COX-2, which are known to promote tumor angiogenesis; F: Single cell measurements of nuclear NF-κB p65 and DNA in the breast cancer tissue shown in B, which shows a subpopulation of cells with nuclear localization of NF-κB p65, promotes breast cancer cell migration and metastasis; G and H: CD68 (green as in panel A), NF-κB p65 (red as in panel B) and DNA (blue) showing stroma-restricted infiltration of CD68+ macrophages (G) and macrophages infiltrating tumor nests (H) (examples indicated by arrows); I: PD-1-expressing tumor cells (red plasma membrane signal) surrounded by CD45RO+ memory lymphocytes (green plasma membrane signal) in the stroma and tumor nests (examples indicated by arrows); J-M: p53 (yellow nuclear signal), Ki-67 (green nuclear signal), pan-cytokeratin (red plasma membrane signal) showing heterogeneity in p53 overexpression within tumor nests (J) and proliferating tumor cells and stromal immune cells, indicated by white and yellow arrows, respectively.

3.3) Integration of digital imaging with other data sets

Multi-region spatial DNA sequencing has revealed significant intratumoral genetic heterogeneity that is underestimated by sequencing of single tumor biopsies. Quantitative pathology with digital imaging of intact tissues can be coupled to other platform technologies to allow high content analysis of various macromolecules and whole genome sequencing in specific tumor microenvironments and cell populations to fully assess intratumoral heterogeneity. Tumor microenvironments and cell populations can be identified by digital imaging and captured by laser capture microscopy (LCM) or by coring of tumor samples for further genomic and proteomic analysis. Sequencing of specific regions or the whole genome can be performed on the purified cell populations. Proteomic profiling by mass spectrometry has been applied to highly enriched captured cell populations to characterize breast tumor microenvironments and to elucidate specific regulatory pathways involved breast tumorigenesis. Profiling of
mutations in captured cell populations has also been used to characterize genetic heterogeneity and to aid cancer diagnostic testing.\textsuperscript{227,229} Gene expression profiling has also been applied to LCM-enriched epithelial cell populations to identify signaling pathways associated with specific subtypes of breast cancer\textsuperscript{230,231} and to elucidate signatures associated with epithelial and stromal compartments that have diagnostic and prognostic significance.\textsuperscript{213,232} DNA methylation analysis in purified cell populations from tumors has revealed changes in methylation of genes in specific cell types within the tumor system.\textsuperscript{233,234}

The integration of the above data sources with digital imaging technology and machine learning can be accomplished by employing a cellular/tissue systems biology approach like that illustrated in Figure 6. Genomics and Proteomics have inferred pathway maps associated with normal tissue and alterations of those pathways associated with diseased tissue. Pathway maps provide insights into the composition and topology of the signaling network, however: they cannot be used to determine activation states of key proteins in a pathway; they do not provide quantitative information on relationships between pathways; and do not provide information about cell-to-cell variability.\textsuperscript{235} Multiparameter imaging of tissues and dynamics of cellular models, along with other platform data, combined with machine learning enables the identification and characterization of the dynamics and heterogeneity in signaling pathways, and the construction of computational models that provide a deeper understanding of the normal and abnormal functioning of those pathways.
3.4) Tools to Address Heterogeneity in Tissue Systems

Digital imaging of multiplexed biomarkers in tissues enables assessment of both genetic and non-genetic heterogeneity in tumor tissue systems. Quantitative multivariate data can be extracted from digital tissue images of protein, DNA and RNA biomarkers at single cell, subcellular and tissue compartment levels and spatial relationships can be measured. The digital imaging challenge is to separate neighboring and overlapping nuclei to enable single cell and subcellular analyses. Approaches such as the watershed algorithm use grayscale information to separate nuclei by finding valleys of low intensity between the high intensities of nuclei centers. A-priori knowledge about the shape, size and intensity distribution of fluorescently-labeled nuclei can also be used with pattern recognition algorithms, to guide image analysis algorithms and improve nuclear segmentation. Once nuclei are segmented the cytoplasm and plasma membranes of each cell can either be estimated based on the nuclear masks or specifically masked using cytoplasmic- and plasma membrane-restricted biomarkers. There are many commercially available image analysis software packages from companies such as Definiens, AG (Munich, Germany), HistoRx (Branford, CN, USA), Hamamatsu Photonics, K.K.
(Hamamatsu City, Japan)\textsuperscript{186} that can segment nuclei and enable quantitative measurements of biomarkers in subcellular compartments and measure tissue morphometric features.

Variation in tumor microenvironment is thought to be an important driver of tumor cell heterogeneity, as well as an impediment to treatment.\textsuperscript{240} In order to understand the relationship between cellular heterogeneity and microenvironment, it is necessary to characterize the local environment of each cell in addition to the biomarker activities. Figure 7 illustrates an algorithm designed to identify, and classify tissue regions with similar microenvironments. The image is divided into a series of small patches (Figure 7A). A feature set that characterizes biomarker intensity distributions within and between cells is measured for each patch. The similarity matrix (Figure 7B) indicates the degree of similarity between patches. The similarity matrix is converted to a network (Figure 7C) that is used to cluster the patches into groups with distinct microenvironment.\textsuperscript{241-243} The correlation of biomarker activity with microenvironment class will be useful for the identification of biomarkers that are more commonly affected by the local environment and aid in the interpretation of the heterogeneity exhibited by those biomarkers.

\textbf{Figure 7. Automated Identification of Cancer Cell Microenvironments.} (A) A fluorescent image is broken into small homogeneous texture regions (“patches”) and a feature vector is measured for each image patch (e.g. distributions of biomarker intensities on nuclei, cells, and cytoplasm). (B) Matrix S
indicates the degree of similarity between each texture patch. Non-zero elements of matrix S are indicated by black dots. (C) The similarity matrix can be converted into a network, so that each image patch is a node in the network, and values in S describe weights on edges between image patches (shown by the thickness of the connecting edge). By performing a random walk on the network, identifying bottlenecks and removing them, a clustering of the nodes into groups of homogeneous image patches emerges naturally- each cluster containing a distinct microenvironment (shown in colored circles in panels B and C).

While computational imaging provides quantitative information on the structure and function of cells and sub-cellular structures, there are several key challenges in addressing tissue heterogeneity in digital tissue images. Robust, scalable image analysis tools are required that can operate in extremely low signal to noise ratio regimes and handle the data intensive challenges presented by multiplexed and hyperplexed tissue biomarkers. Computational approaches must carefully analyze the long-tail behavior of biomarker intensity and feature distributions in order to characterize heterogeneous sub-populations of cells. Furthermore, the reasoning strategies of pathologists need to be incorporated into analysis approaches to flag diagnostically relevant areas of tissue images for cancer detection and analysis. These challenges can be addressed by machine learning approaches, which enable expert domain knowledge and rule-based decision-making to be incorporated to guide tissue image analysis.

4) Computational and Systems Biology Methods in Cancer Research and Diagnostics

4.1) Machine learning and other computational methods in cancer research, drug discovery and diagnostics

In parallel to the dramatic expansion of automated instrumentation for collecting biomedical research data over the past thirty years, machine learning has arisen as a powerful approach to the analysis and interpretation of this and many other kinds of data.\textsuperscript{244} The field of machine learning has already led to many revolutionary technologies, from autonomous vehicles to voice commands to question-answering systems such as Watson, and promises much more. There are
two distinct paradigms for posing questions using machine learning, with many intermediate approaches that combine them. These are *supervised learning*, or learning by example, and *unsupervised learning*, or learning models from data. In both approaches, a large set of objects/observations are provided that are described by various *features*. Objects can include individual cells or patients and features are measures such as cell size or blood antigen levels.

Within supervised learning there are two types of challenges: i) *classification*, i.e. recognition of specific object classes, and ii) *regression* to estimate the value of output variables. In both supervised problems, the class names or output variable values for some of the objects form a *training set* and the task is to learn a rule(s) that allow the class/value to be predicted/estimated from the feature values for new objects. Accuracy of the system is measured using a *test set*, or the training set can be divided into portions and some portions used for training and others for testing.

There are many frameworks for supervised learning, ranging from simple linear approaches (linear discriminants, nearest neighbor classifiers) to piecewise linear (decision trees) to highly non-linear (artificial neural networks, support vector machines with nonlinear kernels). All have been increasingly used in biomedical research, particularly in analyzing results from microarray experiments. For example, gene expression profiles have been used to distinguish tumor samples from normal tissue. Similar approaches have been used to correlate serum protein profiles measured by mass spectrometry with the presence of particular cancer types in patients. Classification of gene expression data can also provide information for determining treatment or prognosis. This approach has been used to distinguish the tissue site of tumor origin and to distinguish metastatic from non-metastatic tumors.
Similar approaches can be used to analyze biomedical images, in which the most challenging task is to decide upon, and calculate, the features that describe a particular image or region of an image. Magnetic resonance imaging (MRI) is widely used for detection and staging of tumors and performance on these tasks can be improved using machine learning. Supervised learning methods have also been applied to analysis of traditional histopathology and have achieved similar, and often superior, results to manual scoring by pathologists. These approaches use intrinsic tissue structural information; however, additional information can be gained by analyzing molecular changes in cancer tissues, such as gene expression levels or subcellular localizations of proteins.

In unsupervised learning, only feature values are provided and analogous to classification and regression in supervised learning, the challenge is either to discover classes or clusters that are present, or to build a model that allows estimation of some of the features from the others. The principle behind cluster analysis is to group observations based on similarity in feature values as measured by a distance function. Cluster analysis has been widely used to find clusters of genes with similar changes in expression level in response to some stimulus. A variation on this theme is to perform cluster analysis and then validate the clusters by finding other distinguishing characteristics. For example, patients have been grouped into clusters based on DNA microarray data, and those clusters were shown to differ in clinical phenotypes. In HCA, cluster analysis can be used to discover patterns that are present in a collection of images whether for learning protein localization patterns, or for grouping compounds by their effects on cells. In addition to characterizing average phenotypes across a population of cells, heterogeneity of responses can be determined automatically by finding clusters of cells that differ in their multivariate phenotype. This approach has been demonstrated to be useful for predicting the
response of various tumor cell lines to chemotherapeutic agents.\textsuperscript{165}

While most machine learning applications use the basic paradigms of supervised and unsupervised learning, there are many intermediate approaches termed semi-supervised learning that use a degree of supervision to accomplish unsupervised learning goals. For example, unlabeled data can be used to learn probability distributions for features and this knowledge can improve classifier training. Alternatively, labeled data can be used to learn a distance function that is then used for unsupervised learning. Human intervention can also be used to tune distance functions or classifiers; this is termed interactive supervision. An initial clustering can be done using all of the data without labels, and the distance function parameters can be optimized based on user feedback. Alternatively, examples of two or more classes can be given with classes assigned for unlabeled data, and then corrections can be made or new classes defined.

Interactive semi-supervised learning can be a form of what is termed \textit{active} machine learning. A semi-supervised learning system would be termed active if the system chooses which points the human should focus on rather than displaying all points and allowing the human to choose points. Most machine learning methods assume that all of the data, i.e. features and/or labels, for a problem have already been obtained. Active learning deals with situations in which the data is incomplete but where it is possible to acquire additional data. Active learning begins by building a predictive model from currently available data using one or more of the methods above. Various methods are then used to decide which missing data to acquire, with the goal of maximally improving the model. Active learning approaches are likely to be highly relevant to HCA as part of drug development.\textsuperscript{80} Selection of drug candidates must be guided both by desired effect on a given target and by lack of effects on other cellular targets that may lead to
side effects. Given that the number of potential targets is over ten thousand, and that the number of potential drugs is at least one million, screening all combinations would be prohibitively expensive. An alternative is to assay a representative subset of targets thought to reflect major cellular pathways.\textsuperscript{259} This is clearly an improvement over current practice, but the complexity and interconnection of those pathways present significant challenges.

Two algorithmic approaches, namely \textit{robust statistics},\textsuperscript{260} and \textit{spectral graph-theory},\textsuperscript{261} from the fields of computer vision and machine learning have the potential to address emerging challenges in digital tissue biomarker analysis. Robust statistics provides a suite of algorithms\textsuperscript{262} to: (i) fit parametric models to noisy data, e.g., to model a membrane wall with a parametric shape and extract the parameters of the shape model from noisy images of a membrane wall; (ii) instantiate multiple occurrences of a model whose number and type must be determined from the data, and (iii) account for structured outliers, i.e. discount certain observations in the data so that they do not affect the estimation of the model parameters. Algorithms from robust statistics and spectral clustering have significantly advanced the state-of-the-art in the fields of computer vision and machine learning. These approaches are expected to be highly relevant to and widely applicable to the computational tasks of analyzing digital pathology datasets.\textsuperscript{263-265}

\textbf{4.2) Computational modeling of cellular systems}

Heterogeneity and complexity of cancer also pose major challenges for computational modeling. Cancer has signatures and mechanisms that operate at multiple scales. Two fundamental scales that are bridged by 2D and 3D cellomic data are the intracellular molecular network scale and the cellular scale at which cell-cell interactions take place. Just as experimental models must go beyond the uniformity of a single cell line, computational models at the molecular level must
consider far more than single networks in isolation if they are to provide a realistic description. At the same time, models of processes at the cell population level must incorporate molecular details if they are to reveal underlying mechanisms through which cancer mutations and other disease-associated changes exert their effects. Bridging between the molecular and cell population levels is a major challenge for computational modeling going forward. Achievement of this integration will lead to improved mechanistic understanding of cancer and holds promise for the development of new therapies.

In a mechanistic model, cellular activity is represented as a directed network of chemical interactions. Each node of the network represents a type of molecule in or around the cell, and each network edge represents a chemical interaction. The system’s behavior is determined by the network topology, molecular concentrations and reaction constants, as illustrated in Figure 8.

**Figure 8. A simple reaction network and its output as an example of mechanistic pathway modeling.** In panel (A), reactants A and B (green) reversibly combine in equal amounts to form the homodimer AB (red), which catalyzes the conversion of C to D (blue). D in turn catalyzes the dissociation of AB into A and B. The system of equations represented by this network can be solved to find how the concentrations evolve in time. Panel B shows the time evolution of molecular concentrations if A, B and C are initially present in equal amounts and AB and D are initially absent. The shapes of the curves depend on the rates of the chemical reactions.
The first major task in computational modeling is to identify a model or set of models that is compatible with both prior knowledge about underlying molecular and cellular mechanisms and the available data. In the discussion below we focus on the specific requirements for developing models based on HCA data, but the procedures for developing models are quite general. Figure 9 illustrates the procedure for constructing a Rule-based model, using the epidermal growth factor receptor (EGFR) pathway as an example. Rule-based modeling languages like BioNetGen\textsuperscript{266} bridge the gap between the design of the model and the mathematical definition of its computation. Mechanistic modeling is distinct from the machine learning approaches in that the underlying models are biophysical and biochemical in origin rather than purely descriptive of the data.\textsuperscript{267,268} The two approaches share a common problem, which is the identification of model parameters that are consistent with a given set of data.\textsuperscript{269} This step is necessary to generate predictions from the model and to estimate the confidence or uncertainty in those predictions,\textsuperscript{270} but is currently a major bottleneck in the development of all forms of mechanistic models. This problem is particularly acute for large-scale models that have potentially tens to hundreds of unknown parameters.\textsuperscript{271,272}
Figure 9. Steps in rule-based modeling of intracellular signaling. Building a pathway model, such as the epidermal growth factor receptor (EGFR) pathway involves: A. Create a map of pathway components and interactions; B. Translate elements of the map into molecules (blue) and rules (red) using, for example, the syntax of the BioNetGen language; C. Estimate the parameters, such as initial concentrations of proteins and rate constants, and calibrate model; and D. Run simulations using one of several methods such as ordinary differential equations (ODE), stochastic simulation algorithms (SSA), or network-free simulation (NFsim). Steps C and D are frequently iterated as model predictions are used to drive experimental studies and the resulting data are used to refine and recalibrate the model.

Here we identify four main challenges for developing computational models that integrate HCS data: network complexity, cell-to-cell variability, spatial complexity, and multi-cellularity, which we consider in the following paragraphs.

Network complexity. Mechanistic models of cell signaling are usually based on standard chemical kinetics descriptions that arise from treating the cell as a well-stirred chemical reactor. Such reaction network models of cellular processes can be simulated by numerical integration of differential equations (if the cellular concentrations are large enough to be approximated as
continuous) or through kinetic Monte Carlo simulations (most commonly, the Gillespie algorithm\textsuperscript{273}), if the noise arising from small molecular concentrations is important. This approach is typically used to model systems ranging in size from a few species to as many as a few hundred, although parameter estimation becomes problematic for models of such complexity. More coarse-grained simulation approaches, such as Boolean network\textsuperscript{274} or fuzzy-logic approaches,\textsuperscript{235} which may reduce the problem of parameter estimation, have also been used as a basis for automated methods of learning model structure, i.e., the rules that govern interactions among the model components. Recently, statistical methods for model selection have also been applied to reaction network models to determine model architectures most compatible with experimental data.\textsuperscript{275}

A major problem in cancer modeling is to predict the effect of observed mutations on the network response. As mutations tend to effect specific molecular interactions, achieving this predictive capability drives the development of increasingly complex models. Unfortunately, the reaction network approach suffers from the problem of combinatorial complexity, which results in an explosion of network size – the numbers of species and reactions that must be considered in the model – as the complexity of molecules and number of interactions grows. This problem makes manual specification of reaction networks, the standard approach, time-consuming and ultimately prohibitive. The cost of simulations also grows rapidly with network size and can become prohibitive even for relatively small systems.\textsuperscript{276} The recent development of rule-based modeling\textsuperscript{277,278} largely eliminates the effects of combinatorial complexity in the specification and simulation of models, although the general issue of calibrating the parameters of large-scale biological models remains open.
Cell-to-cell variability. In principle the effects of cell-to-cell variability in protein expression and other basic parameters such as volume or surface area can be modeled using reaction network models. For example, the distribution of responses to a fixed stimulation of a cell population could be modeled if the distributions of the key signaling components affecting the response were known. Recently, protein expression levels have been measured on a genome scale in single cells and shown to largely follow simple distributions characterized by only a few moments, e.g., mean, variance, and skewness. If the necessary antibodies are available, HCA data can also be used to parameterize expression level distributions for a cell population, and these parameters can be fed into a model to predict the distribution of responses. An important additional issue, that has been addressed very little, is how correlations in protein expression affect variability. In one study of T cell signaling, positive correlation between expression levels of activating and inhibitory components of a signaling pathway was shown to reduce the effect of cell-to-cell variations in those proteins. A number of studies have shown that variability in protein expression can lead to dramatic variation in cellular responses, leading to bimodal distributions of activity or outcome. Such effects typically arise from nonlinear effects such as positive feedback, but can have other origins. Cell populations may use such variability as a form of bet-hedging to avoid making costly decisions, such as whether to grow or undergo apoptosis, in response to possibly spurious signals. In the context of cancer, the effect of such variability complicates the development of effective treatments.

Spatial complexity. Imaging provides a wealth of information about the spatial localization of molecules within cells. This localization is critical to many signaling processes. The movement of molecules between cell compartments is a major component of information flow and signal processing, e.g., endocytosis and transport of complexes through nuclear pore complexes. A
range of modeling and computational approaches have been developed to describe such
effects. At the simplest level, compartments can be introduced to reaction network models to
prevent interactions between components in different compartments. Models based on partial
differential equations [VCell] or Brownian Dynamics [MCell; Smoldyn; GFRD] handle spatial effects at a much higher level of detail. Depending on the level of resolution that is chosen, inclusion of spatial effects increases the simulation cost and the difficulty of fitting model parameters. Currently these costs preclude the possibility of model selection in most cases. At the same time, software for spatial simulation is advancing rapidly and can be expected to play a more important role as the amount of HCA data increases.

**Multicellularity.** The most daunting challenge for modeling HCA data comes in handling cell-cell interactions and coupling those to the molecular scale. Cell behaviors are coupled through the secretion and uptake of ligands, competition for resources, and through direct cell-cell contacts. Modeling these effects requires multi-level models that couple intracellular regulatory networks to mechanisms for cell-cell communication. Cells are typically modeled as agents in such models. In order to develop multi-level models, standard methods for modeling cells as discrete agents in a population must be combined with reaction network modeling methods. Currently there are relatively few software tools available for the construction of such models, and most of the models that have been developed use special-purpose code. Use of multi-level modeling has grown in the past few years and has been accompanied by the development of several new frameworks for developing multi-level models, e.g. ML-rules; chaste; etc. The major issues going forward will be the computational expense of simulating such models and the accompanying issues of model parameterization and model selection.
Recently, the first truly comprehensive model of cellular processes and their regulation was developed for mycoplasma genitalium, one of the simplest known organisms. The model demonstrates that a wide range of cell processes can be modeled and integrated into a single computational model, but the effort required to construct and simulate the current model, which describes the behavior of only a single cell under controlled conditions, demonstrates the magnitude of the challenges that lie ahead.

**Conclusion and Outlook**

The integration of all imaging modalities with genomics and proteomics will have an important impact on our understanding of the molecular basis of cancer. In particular, HCA coupled to computational biology will yield the necessary statistical analyses to define and understand the impact of heterogeneity in both diagnostics and therapeutics. Although challenging, the development and application of systems biology tools should make it feasible to begin modeling the complexities of cancer.

**Acknowledgements**

This work was supported in part from research grants from the State of PA-Tobacco Funds (DLT), NIH R01 GM075205IH (RFM), NIH-NIGMS R01 GM086238 (TL, Ivet Bahar, PI), NIH-NCRR UL1RR024153-05 (TL and JRF, Steve Reiss, PI) and NIH-NCI 1R21CA164433-01A1 (CC, Liu, PI), Pennsylvania Department of Health Cure Program Grant RFA#10-07-03 (RC-T). The authors thank Andy Stern for discussions on cancer biology and drug discovery, Adrian Lee and Steffi Oesterreich for discussions on cancer biomarkers, Ivet Bahar for discussions on computational biology, Jennifer Grandis for insights in head and neck cancer and
Cal33 cells used in creating Figure 3 and Kumiko Isse and Anothony J. Demetris for supplying the images used for the analysis shown in Figure 7.

References


41. Giacomotto J, Ségalat L. High-throughput screening and small animal models, where are we? Brit J Pharmacol 2010; 160: 204-16.


