Collaborative Analysis of High-Content Image Data

Vijay S. Kumar, Jenny Weisenberg Williams,
Kareem S. Aggour, Brion Sarachan
Software Sciences & Analytics, GE Global Research
1 Research Circle, Niskayuna, NY 12309, USA
{v.kumar1, weisenje, aggour, sarachan}@ge.com

Yousef Al-Kofahi, Alberto Santamaria-Pang
Diagnostics, Imaging & Biomedical Technologies
GE Global Research
1 Research Circle, Niskayuna, NY 12309, USA
{alkofahi, santamar}@ge.com

Abstract—**In this paper, we show how popular open-source data analytics platforms, such as KNIME, coupled with industry standard cluster computing systems such as Hadoop and Spark can be leveraged to build a highly flexible, scalable and user-friendly collaborative framework for the analysis of high-content hyperplexed molecular imaging data in a public cloud.**

Index Terms—**Large-scale Imaging, Molecular Pathology, Hyperplexed Microscopy, KNIME workflow, Big Data systems, Collaborative framework, Amazon Web Services Cloud**

I. Introduction

In the era of precision medicine, deep analysis of image data acquired from advanced microscopy imaging instruments is critical to life science research and improved healthcare. The application of biomedical imaging informatics techniques to molecular pathology data enables the precise characterization of tissue specimens at cellular and sub-cellular granularities. Biologists and other life science researchers rely heavily on these characterizations to identify and develop novel diagnostic methods for complex disease conditions such as cancers.

Molecular pathology, the discipline which couples molecular information such as protein expression with tissue morphology, has been greatly expanded over the past decade through the adoption of fluorescent technologies that allow for multiplex staining to probe for the presence of biomarkers in tissue samples. The spatial distributions and signal intensities of markers within these samples represent the levels of expression of proteins within different regions of the tissue. The maximum number of distinct biomarkers that can be examined at once in a given sample is limited by the number of distinct fluorescent channels that can be imaged – typically, this number is less than 5. In extreme cases, the use of hyperspectral imaging can increase this number to about 10.

GE Global Research has developed a novel hyperplexed microscopy imaging technique called MultiOmyx™(*) for sequential staining, with iterative chemical removal of stains between rounds, such that the same tissue specimen can be repeatedly re-stained with a larger number (~60, possibly more) of different antibodies [1]. The MultiOmyx technology coupled with novel single cell analysis can generate a wealth of spatially-distributed biomarker information from a single physical tissue sample. This in-situ analysis of multiple proteins in a single sample can provide a comprehensive molecular profile for the specimen in question, thereby enabling biologists to carry out large-scale biopharmaceutical research studies targeting precision diagnostics.

When combined with high-throughput tissue analysis methodologies such as Tissue Microarrays (TMAs), hyperplexed microscopy can generate unprecedented volumes of high-content image data. For instance, a typical cancer study could comprise samples from a cohort of 300 subjects with 3 TMA *cores* per subject (typically 0.6mm in diameter/core). In tumor samples, we routinely find 1,500 to 3,000 cells per FOV and 40-60 markers imaged producing expression values for 3 cellular compartments (nucleus, cytoplasm, and cell membrane). An enormous amount of quantitative information (hundreds of millions of data points) is generated per study which, when correlated against known clinical information, can enable discovery of new biomarkers for improved disease diagnosis and therapeutic selection. As manual review of such large, content-rich data is prohibitive, automated computational approaches are of paramount importance. The availability of novel algorithms to analyze hyperplexed data notwithstanding, researchers often complain about the lack of a software framework that will allow them to efficiently transform large image data into a custom set of quantitative features for exploratory correlation hypothesis testing and visualization.

Working closely with GE cancer biologists, we gathered the following set of key requirements for a software framework to enable MultiOmyx image and data analysis:

1. Increased Productivity: Currently, researchers spend days or even weeks extracting biomarker information from raw image data. Speeding up this process will enable them to instead spend majority of their time performing statistical analysis on processed data to discover novel biomarkers.
2. Flexibility and Ease of use: Researchers want a self-service, plug-and-play style analysis model that is easy to understand, and allows them to perform custom data analysis operations themselves, without the need to rely on software experts or statisticians for each experimental run.
3. Reliability: Researchers want a computational platform capable of generating reproducible analysis results even in the presence of unexpected hardware/software failures.
4. Secure Collaboration: Researchers want the ability to share datasets, analysis, and findings with collaborators from external organizations in a secure, timely manner.

Additionally, researchers also want to avoid vendor lock-in and expensive license/subscriptions associated with commercial software. This paper describes how we developed a software framework to address the above requirements exclusively using

[*] Trademark of General Electric Company

open-source solutions. Specifically, we describe our prototype of such a framework built using a popular analytics platform called KNIME [2] and industry-standard cluster computing systems like Hadoop [3] and Spark [4]. We also demonstrate its use in enabling collaborative analysis for MultiOmyx in the Amazon Web Services (AWS) cloud.

II. HIGH-CONTENT IMAGE DATA ANALYSIS

Tissue specimen content is structurally complex. Tumor samples, for instance, contain a multitude of cell and tissue types including epithelial, stromal, smooth muscle, nerve and immune cells. Cells, in turn, consist of nuclei, membrane and cytoplasmic components. The primary goal of MultiOmyx analysis is the in-depth examination of protein expression values within these tissue and cell structures based on average signal intensity, expression ratios between cell compartments or other statistical determinations of expression such as standard deviation across cell populations. Using the relative distribution of one or more biomarkers within tissue regions of interest, biologists may test multiple hypotheses (e.g., Do the number and sub-type of T-cells associated with a primary tumor predict therapeutic response? Or, is localization of the cells the critical predictive feature?). Given the large number of markers imaged in MultiOmyx studies, the number of such hypotheses could grow exponentially. Analysis of high-content image datasets is carried out using well-defined pipelines of data processing operations [5]. Figure 1 shows a high-level pipeline of operations for analyzing MultiOmyx image data.

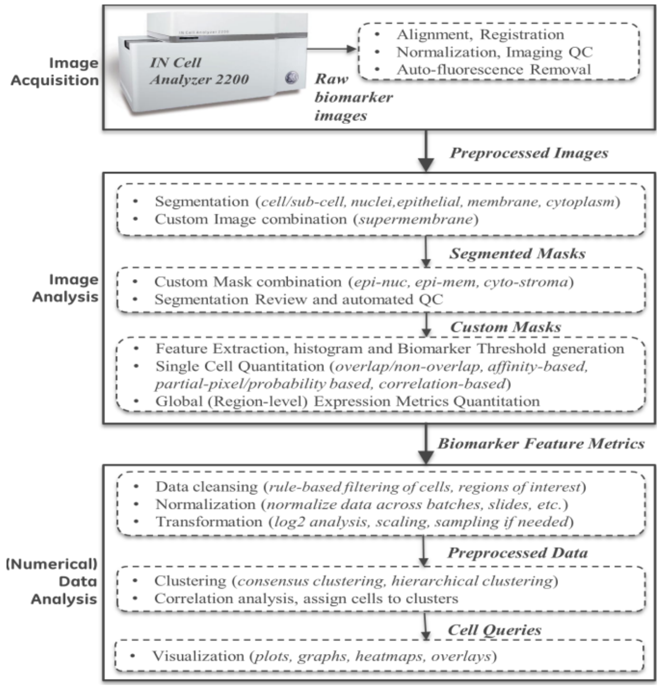


Image Acquisition
IN Cell Analyzer 2200
Raw biomarker images
Alignment, Registration
Normalization, Imaging QC
Auto-fluorescence Removal
Preprocessed Images
Image Analysis
Segmentation (cell/sub-cell, nuclei,epithelial, membrane, cytoplasm)
Custom Image combination (supermembrane)
Segmented Masks
Custom Mask combination (epi-nuc, epi-mem, cyto-stroma)
Segmentation Review and automated QC
Custom Masks
Feature Extraction, histogram and Biomarker Threshold generation
Single Cell Quantitation (overlap/non-overlap, affinity-based, partial-pixel/probability based, correlation-based)
Global (Region-level) Expression Metrics Quantitation
Biomarker Feature Metrics
(Numerical) Data Analysis
Data cleansing (rule-based filtering of cells, regions of interest)
Normalization (normalize data across batches, slides, etc.)
Transformation (log2 analysis, scaling, sampling if needed)
Preprocessed Data
Clustering (consensus clustering, hierarchical clustering)
Correlation analysis, assign cells to clusters
Cell Queries
Visualization (plots, graphs, heatmaps, overlays)

Fig. 1. Canonical End-to-end Data Analysis Pipeline in MultiOmyx

Depending on the specific study, the nature of the disease and cell types involved, biologists assemble custom analysis pipelines using combinations of some or all these operations. The outputs from these pipelines may be used for patient stratification, studying co-localization patterns, differential diagnoses and validating markers on new patient cohorts.

III. RELATED WORK

Existing well-established image processing libraries like ImageJ/Fiji [6] provide interfaces for easily plugging in custom biomedical image analysis algorithms. However, ImageJ is not designed to process large-scale data, and does not support flexible pipelines. Frameworks like HIPI [7] and OpenIMAJ/Image Terrier [8] use Hadoop to scale up very large-scale image analysis operations, but do not focus on high-content biomedical image data. Bajcsy et al. [9] use Hadoop to process large-scale microscopy image data on compute clusters, but do not provide a framework for analysis pipeline composition. CellProfiler (and Analyst) [10] enables assembly of image analysis pipelines for processing cellular image data. Analogous to high-throughput processing of TMAs, it can efficiently process tens of thousands of images obtained from well plates. Unlike CellProfiler, we use Hadoop and Spark for reliable computation which allows us to leverage on-demand compute clusters on public cloud environments like Amazon's AWS cloud for collaborative research. Finally, newer cloud image processing frameworks like 4Quant [11] and NeCTAR [12] support collaborative analysis orchestration and execution in public clouds, and share similar motivations to our work. However, they are not optimized for analysis of many biomarkers in hyperplexed image datasets. To our knowledge, there is no single open-source framework that supports all the requirements for MultiOmyx analysis listed in Section I.

IV. FRAMEWORK FOR MULTIOMYX ANALYSIS

This section describes our prototypical software framework to support collaborative analysis of large-scale image data in MultiOmyx. Our prototype has been used to support multiple GE-internal cancer research studies, and is also being piloted in cloud-based collaborative efforts with two external research organizations. Salient features of our framework include:

- Loose integration of KNIME [2] with Hadoop [3] and Spark [4] to enable user-friendly composition of pipelines of image analysis operations and their subsequent execution on large compute clusters.
- Standardized model of data exchange between analysis steps in a KNIME workflow to enable highly flexible plug-and-play analysis of MultiOmyx image data.
- Growing library of highly parallelized image and numerical data analysis operations implemented using the MapReduce programming model [13] so that large-scale TMA studies run 45-95 times faster on clusters.
- Seamless transition between local and cloud execution modes, both controlled via a common user interface. Analysis pipelines that run locally on 'on-premise' clusters can be further scaled up by running them on on-demand elastic clusters in the Amazon AWS cloud. The cloud execution mode also supports highly secure collaboration with external researchers.

Figure 2 shows the high-level architecture of our framework and its various components.

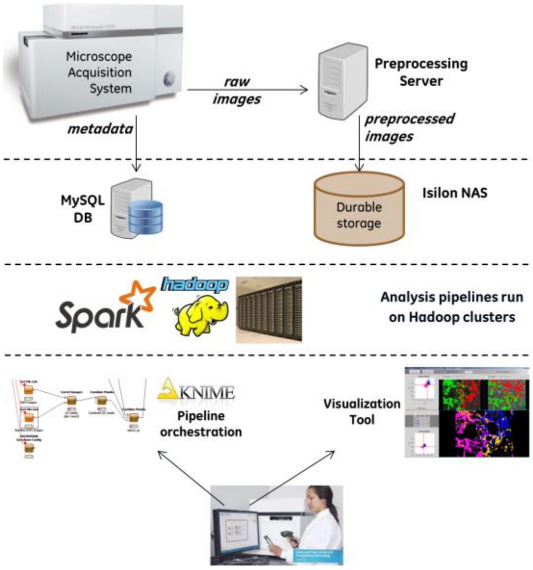
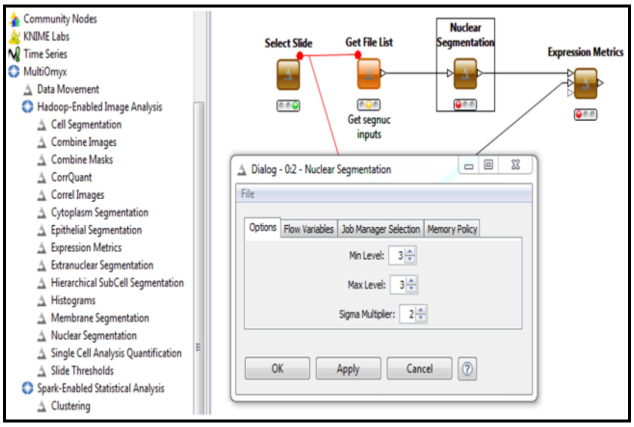


Fig. 2. High-level Architecture of Framework for MultiOmyx analysis

Preprocessed images for all studies are persisted in a durable storage system, while metadata containing detailed information on the acquisition process for each study is captured in a relational database. Biologists interact with MultiOmyx data in two ways – they 1) apply analysis algorithms to process the data, and 2) view images using custom visualization tools.

Our framework uses the open-source KNIME (Konstanz Information Miner) platform [2] to allow biologists to create, configure and execute ad hoc analysis workflows. KNIME has a graphical user interface that allows users to drag-and-drop nodes onto an editor to create such workflows. We created KNIME extensions that allow MultiOmyx–specific analysis algorithms to be plugged in as nodes in a workflow, as shown in fig. 3. In our framework, any algorithm-specific parameters (such as min and max levels for nuclear segmentation), that allow users to adapt algorithms to different image scenarios can be configured within the corresponding node's dialog box.

Additionally, biologists can greatly benefit from flexible pipeline reuse – that is, the ability to take an existing workflow originally created for one study, and re-use it for an entirely different study, by swapping out some workflow nodes and replacing them with relevant ones for the new study. Also, it may be desirable to compare alternative workflows that test the same hypothesis – one may segment the membrane, while the other may examine the cytoplasm instead. In such cases, it makes sense to have a plug-and-play analysis model (e.g., given a list of images, one can apply any segmentation algorithm), instead of creating different workflow instances for each alternative. To facilitate such flexibility, we developed a standardized, self-describing model of data exchange between MultiOmyx analysis nodes by extending KNIME's BufferedDataTable structure. Thus, biologists can avail of self-service, reusable analysis workflows without relying on software experts. Once workflows are created using KNIME, our framework leverages cluster computing platforms, when available, to efficiently execute individual analysis operations.



Min Level: 3
Max Level: 3
Sigma Multiplier: 2

Fig. 3. KNIME extensions for MultiOmyx analysis orchestration

We have accomplished a loosely-coupled integration of KNIME with scalable data processing systems like Apache Hadoop [3] and Spark [4] to reliably process large image datasets on low-cost, commodity hardware. As these systems are not designed to work out of the box with image data, we implemented a parallelized library of MultiOmyx-specific image analysis operations using the MapReduce programming model [13]. By design, each image analysis node runs as a Hadoop MapReduce job, while Spark is used to scale up computationally intensive statistical analysis steps such as data clustering. Our Hadoop jobs exploit the inherent parallelism present at multiple data granularities to decompose analysis into a large number of constituent tasks. We parallelize across studies, across slides in each study, across TMA samples in each slide, and across biomarkers within each sample. However, all the complexity in execution on these cluster computing platforms is transparent to the biologists – our framework takes care of all data movement and job scheduling under the hood.

The most tangible benefit of this framework feature is the reduced analysis execution times. We transitioned MultiOmyx lung cancer analysis of 32 biomarkers and 3000+ samples from a single (12-core) server implementation that the biologists used previously, to our framework running on a 47-server Hadoop cluster. As seen in table I, biologists previously spent almost 2 days every time they ran an analysis workflow that computed both image-level and single cell metrics for this data.

TABLE I. PERFORMANCE IMPROVEMENT FOR PILOT MULTIOMYX STUDY

Slide ID	# of TMA samples	Total # of images	Total image data size	Analysis Execution Time		
				Single server (min.)	Hadoop (47 servers)	
					Parallel by sample	Parallel by slide & sample
1	572	90,978	577.5 GB	~ 480	15 min	66 min
2	516	81,558	508 GB	~ 480	14 min	66 min
3	444	70,182	441 GB	~ 480	12 min	65 min
4	572	90,978	578.3 GB	~ 480	15 min	67 min
5	516	81,558	509.2 GB	~ 480	14 min	65 min
6	444	70,182	438.7 GB	~ 480	12 min	64 min
Total	**3,064**	**485,436**	**3.053 TB**	**48 hrs**	**∑=1.4 hrs**	**max=1.1 hrs**

Our framework ran the same workflow almost 36 times faster when analysis was parallelized only across samples within each slide. When all 6 slides were processed concurrently, execution was 43 times faster than on the single server. This order-of-magnitude improved execution time allows biologists to now spend the majority of their time on exploratory analysis of the metrics and hypothesis testing. More significantly, it leads to newer research methodologies, wherein biologists are able to run many more analyses in a given time than was previously possible. This allows fine-tuning of algorithm parameters and increases the analysis precision and confidence in the diagnosis results. Using our framework, biologists can not only run image analysis operations but also run statistical analysis on the resulting data, either via R code snippets or by using Spark libraries, all within the same KNIME workflow.

V. Cloud-based collaboration for MultiOmyx

Our framework for MultiOmyx analysis supports a cloud execution mode that is currently being piloted in the Amazon Web Services (AWS) cloud to support collaboration with research groups external to GE. As seen in fig. 4, the high-level architecture of our framework's cloud version has some key differences: All MultiOmyx data is stored in the cloud (in encrypted form, for data protection), and is now processed using computing resources provisioned on-demand within the cloud. To this end, our framework integrates a number of managed AWS services including the Simple Storage Service (S3), the Relational Database Service (RDS), and the Elastic MapReduce (EMR) service, with our own services for secure, high-speed upload and download of datasets to and from S3.

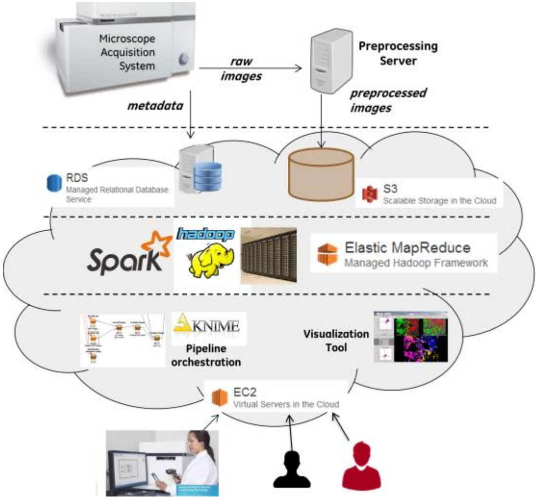


Fig. 4. Cloud-based Architecture for MultiOmyx collaboration

Biologists at GE and external research collaborators access (via remote desktop) virtual workstations in AWS that are pre-installed with the desired analysis and visualization tools. They can collaborate on common workflows via shared installations of KNIME. The KNIME-based interface to data analysis is common across both local and cloud execution modes, making it possible for locally created workflows to be transitioned with minimal changes to run in the cloud. This interface masks users from our complex re-architecture of the framework to execute KNIME workflows in AWS. To save resource usage costs in AWS, analysis here is carried out on transient EMR Hadoop clusters than can be right-sized and provisioned on-demand at the click of a button only for the duration of the analysis. Once analysis is complete, the cluster is shut down. Data is decrypted on the fly as it streams from S3 to EMR clusters for analysis. Likewise, analysis results are encrypted on the fly before they get stored in S3. For each distinct set of collaborators, we currently set up 1) a secure bucket in S3 that stores all data relevant to that collaboration, and 2) a Virtual Private Cloud (VPC) – a secure, isolated sandbox environment to facilitate data analysis for that collaboration. The virtual workstation and EMR clusters for each collaborator are provisioned within the respective VPCs.

Acknowledgment

This research was supported in part by the NIH under the Single Cell Analysis Program Award #RO1CA174377, and by John Burczak under GE's Advanced Technologies program. We would like to acknowledge Scott Pearson and Bill Kupiec for setting up a secure cloud infrastructure, and Fiona Ginty, Chris Sevinsky, Michael Gerdes and other GE biologists for their insights into MultiOmyx data and analysis pipelines.

References

[1] M. J. Gerdes, et al., "Highly Multiplexed Single-cell Analysis of Formalin-fixed Paraffin-Embedded Cancer Tissue", Proc. of the Natl. Academy of Sciences (PNAS) 2013, 110(29): 11982-87.

[2] M. Berthold, et al., "KNIME: The Konstanz Information Miner", in Studies in Classification, Data Analysis and Knowledge Organization, Springer 2008, pp. 319-326.

[3] T. White, "Hadoop: The Definitive Guide", O'Reilly Media Inc., 4th Edition, March 2015.

[4] M. Zaharia, et al., "Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-Memory Cluster Computing",in Proc. of Networked Systems Design & Impl. (NSDI), 2012, pp. 15-28.

[5] J. Rittscher, R. Machiraju and S. Wong, "Microscopic Image Analysis for Life Science Applications", Artech House, 2008.

[6] J. Schindelin, et al., "Fiji: an open-source platform for biological-image analysis", Nature methods 9(7): 676-682, 2012

[7] C. Sweeney, L. Liu, S. Arietta, and J. Lawrence, "HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks", Undergrad. Thesis, Univ. of Virginia, 2011.

[8] J. S. Hare, S. Samangooei, and P. Lewis, "Practical Scalable Image Analysis and Indexing using Hadoop", Multimedia Tools Appl., vol. 71(3), pp. 1215-1248, August 2014.

[9] P. Bajcsy, A. Vandecreme, J. Amelot, P. Nguyen, J. Chalfoun, and M. Brady, "Terabyte Size Image Computations on Hadoop Cluster Platforms", Proc. of IEEE Intl. Conf. on Big Data, 2013.

[10] M.R. Lamprecht, D.M. Sabatini, A.E. Carpenter, "CellProfiler: Free, Versatile Software for Automated Biological Image Analysis", Biotechniques vol. 42(1), pp. 71-75, 2007.

[11] K. Mader, "Interactive Scientific Image Analysis and Analytics using Spark", (presentation), Spark Summit East, March 2015.

[12] T. Bednarz, et al., "Cloud-based Toolbox for Image Analysis, Processing and Reconstruction Tasks", Signal & Image Analysis for Biomedical and Life Sci., Springer 2015, v823, pp. 191-205.

[13] J. Dean, S.Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Commun. ACM, 51(1): 107-113, 2008