This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

BioCyBig: A Cyberphysical System for Integrative Microfluidics-Driven Analysis of Genomic Association Studies*

Mohamed Ibrahim, *Student Member, IEEE,* Krishnendu Chakrabarty, *Fellow, IEEE,* and Jun Zeng, *Senior Member, IEEE*

Abstract—This paper presents a research vision to design a large-scale cyberphysical systems (CPS) experimental framework to enable collaborative and coordinated molecular biology studies. This framework will be based on the integration of CPS with microfluidic biochips and cloud computing. It has the potential to drastically advance personalized medicine through knowledge fusion among many research groups, and synchronization of research planning. This framework therefore leads to a better understanding of diseases such as cancer, and helps researchers in identifying effective treatments. A case study from cancer research is discussed to explain the significance of our framework in promoting coordinated genomic studies.

Index Terms—microfluidics, big-data, genomics, cyberphysical, integration, Software-as-a-Service, Internet-of-Things, fusion

1 Introduction

PERSONALIZED medicine represents a bold research effort that can revolutionize healthcare and the treatment of diseases such as cancer. In the past, most medical treatments were designed for the "average patient"; such an approach, known as the one-size-fits-all, was deemed to be successful for some cancer patients but not for others. As a result, the Precision Medicine Initiative was launched in 2015 with a $215 million investment to provide clinicians with new technologies, tools, knowledge, and match therapies to patients [1]. Advances in precision medicine will lead to powerful new techniques and treatments that are tailored to specific characteristics of individuals, such as a person's genetic makeup.

Microfluidics is a key technology that enables advances in personalized medicine. Breakthroughs in microfluidics and genome technologies can significantly advance personalized cancer treatment and transform clinical diagnostics from the bench to the bedside. Ultimately, with portable microfluidic devices, patients with breast, lung, and colorectal cancers, for instance, will be able to routinely undergo point-of-care molecular testing as part of patient care, enabling physicians to precisely select treatments that improve the chances of cure. These repeated test results, coupled with timestamps, situational information (time, location, and environment of such tests), and personal information (age, weight, height, gender, etc.), will form the data fabric that can not only highlight the medical condition of an individual but also collectively inform the evolution of the state of the population, for instance, early detection of potential outbreak of highly contagious diseases. In other words, this microfluidic service has the potential to perform complicated genomic studies (e.g., epigenetics [2]) to assist in cancer research, provide point-of-care clinical diagnostics, and accelerate drug development.

One of the most important application areas in cancer-genomics research is the interplay between single-cell biology and omics technologies (i.e., DNA sequences, RNA expression levels, proteomics, and other epigenetic markers) and its impact on disease development and evolution (i.e., association studies) [3]. Single-cell biology utilizes several microfluidic and computational techniques, with an ultimate goal of constructing a genome-wide catalog of genetic, epigenetic, and transcriptomic elements [4], [5]. As shown in Fig. 1, single-cell microfluidic techniques are used to generate omic data from cancer cells; this data is managed and analyzed by computational methods to identify clusters, lineages, and networks, which in turn generate new biological hypotheses. In other words, the contributions to data analysis include two aspects: (1) improved, large-scale machine learning techniques through big genomic data, which will result in more powerful algorithms; (2) human-centric collaborative environment to facilitate communication, collaboration, and synchronization of diverse, microfluidics-based research facilities. Next, biological findings, in turn, guide the development of new microfluidic experiments and computational studies [6]. Specifically, the circular process shown in Fig. 1 needs to be selectively iterated hundreds to thousands of times using cells from a variety of cell populations and tissues. Such an iterative approach enables us to draw precise conclusions about the types and states of these cells, the effective biomarkers that influence genomic or transcriptomic behavior at different loci, and to finally re-focus the scope of subsequent iterations of single-cell analysis. Nevertheless, the application of genome-wide association studies at the single-cell level is extremely complicated and it is hindered by several technical limitations, one of which is the need for a reliable, self-adaptive, and high-throughput control scheme that readily coordinates the experiments of single-cell analysis at a large scale.

Current research methods for single-cell genomic-association studies belong to one of the following categories: (1) extrapolated single-cell methods, which rely on the in-vivo findings of a certain level of biological systems or a single omic data type, such as DNA sequences or RNA expression levels, to extrapolate the results or conclusions of cell subpopulations; this approach is

**This work was supported in part by the National Science Foundation under grant CNS-1135853.*

M. Ibrahim and K. Chakrabarty are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA (e-mail: mohamed.s.ibrahim@duke.edu; krish@ee.duke.edu). J. Zeng is with HP Labs, HP Inc., Palo Alto, CA 94304, USA (e-mail: jun.zeng@hp.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

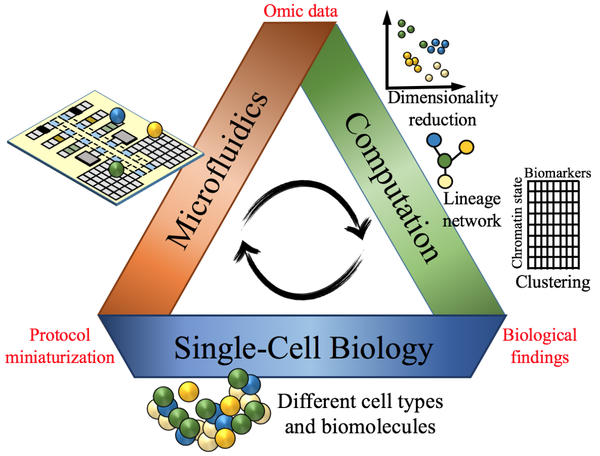

Fig. 1: Iterative genomic-association analysis using microfluidic and computational methods.

pursued by experimentalists using high-throughput microfluidics on real biological systems; however, the analysis that assesses the variation of only a single omic data type can miss complex models that require variation across multiple levels of biological regulation, (2) static data integration techniques, which are applied to large omics data generated from multi-source experimental setting [3]. The development of such analytical methods aims to harness the utility of these comprehensive high-throughput data to elucidate important biomakers. Nevertheless, decoupling such analytical methods from experimental biology expertise does not lead to efficient search techniques for patterns over very large collections of omic data in very high dimension [7]; i.e., raw massive genomic data is not efficiently exploited.

It is evident that the gap between the two approaches represent a technological barrier for researchers against developing an integrative, highly adaptive analysis framework that can relate changes in molecular measurements to disease development, behaviour, and evolution. We describe our vision to close this gap by investigating an interactive CPS for a cloud-based microfluidic service in the Internet of Things (IoT) framework, referred to as BioCyBig. This CPS framework will introduce Microfluidics-as-a-Service (MaaS) for genomic association studies and provide the following benefits:

- It will coordinate the operation of a large number of microfluidic devices (referred to as nodes) to dynamically process iterations of single-cell analysis with high-throughput sequencing control. This approach will pave the way for IoT-enabled real-time collaborative experiments, whereby large number of labs and researchers will be able to coordinate experiments, guide each other, and immediately make decisions on follow-up biochemical protocols or procedures.
- It will leverage the capability of a big-data infrastructure to cumulatively build and enhance the accuracy of biomarker-influence and lineage networks besides cell-type clustering.
- By coupling cyberphysical integration and big-data infrastructure, it will introduce a physical-aware (self-adaptive) microfluidic system, which can reconfigures its nodes (i.e., refocus its analysis scope) based on the dynamic restructuring of computational models—such reconfiguration can be performed either automatically or using human-in-the-loop. This design facilitates sharing genomes and related omics data among researchers, and enables the coordination of thousands of nodes.

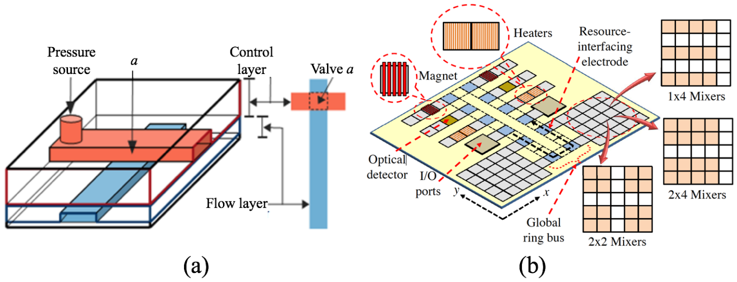

Fig. 2: Schematic view of: (a) flow-based microfluidic biochips; (b) digital-microfluidic biochips.

This paper presents a framework to investigate the interplay between different technologies—microfluidic biochips, biochemical analysis protocols, cyberphysical adaptation, as well big data and cloud computing. We have identified a problem that currently has no solution: a large-scale biochemistry experimental framework based on microfluidics as a service. The proposed multi-layer system architecture and control mechanisms would allow an efficient distributed experimental infrastructure to blossom.

The rest of the paper is organized as follows. An overview about microfluidics technology and biological systems is introduced in Section 2. A case study from cancer research is presented in Section 3 to demonstrate the significance of the proposed framework, BioCyBig. Next, Section 4 depicts the overall architecture of BioCyBig, whereas Sections 5, 6, and 7 discuss the anticipated design challenges and research opportunities at the application level, the microfluidics level, and the middleware level, respectively. Finally, conclusions are drawn in Section 8.

2 Background

In this section, we present an overview of microfluidics technologies. In addition, the biological pathway of gene expression is also elucidated.

2.1 Microfluidics Technology Platforms

Microfluidic biochips (or lab-on-a-chip “LoC”) are typically centimeter-sized devices, with on-chip components having micrometer feature lengths. Miniaturization speeds up chemical reactions and analytical detection; automation and parallelization make it possible to carry out a massive number of different tests simultaneously. These characteristics, especially the delivery of results for a large number of tests within a short amount of time, are especially relevant for clinical diagnostics, genomics, and drug discovery.

Flow-based microfluidic biochips constitute an exciting emerging technology that enables the integration of fluid-handling operations [8]. Continuous liquid flow with picoliter volumes in a flow-based microfluidic biochip can be achieved in etched microchannels in the “flow layer”. Through thousands of integrated microvalves in the “control layer”, different fluid-handling operations, such as mixing, dilution and transportation, can be easily implemented [9], [10] (Fig. 2(a)).

Recent advances in fabrication techniques, including the application of polydimethylsiloxane (PDMS) and dense integration of active microvalves, have enabled the development of flow-based microfluidic biochips. These devices allow a transition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

from a simple topology with only a few channels to large-scale networks of channels for realistic applications [11]. Increasing integration levels provide biochips with tremendous potential; hundreds of different bioassays, i.e., protocols for biochemistry, can be processed independently, simultaneously, and automatically on a coin-sized microfluidic platform [12]. These advances therefore allow massively parallel biochemical processing and immediate point-of-care disease diagnosis [13]. In 2011, Fluidigm, a biotech company that focuses on flow-based microfluidic biochips, launched its initial public offering at NASDAQ, which is a significant milestone in the maturation of the microfluidics industry.

Digital microfluidics has heralded the second (and remarkably advanced) generation of biochips. It is based on electrowetting-on-dielectric (EWOD), which refers to the modulation of the interfacial tension between a liquid and a solid electrode coated with a dielectric layer by applying an electric field [14]. This approach utilizes tiny droplets as on-chip biochemistry carriers. An on-chip array of electrodes, which are addressable through electronic control, can manipulate each droplet electrically, as shown in Fig. 2(b). A set of programmable instructions enables on-chip chemical reactions. Multi-step and complex analytical tasks can be performed with digital microfluidics via a combination of droplet operations (formation, merge, split and migration). By exploiting the reconfigurability inherent in digital microfluidics, these devices are revolutionizing a wide range of applications, such as high-throughput sequencing, parallel immunoassays, clinical diagnostics, DNA sequencing, and protein crystallization.

According to a recent announcement by Illumina, a market leader in DNA sequencing, digital microfluidics has been transitioned to the marketplace for sample preparation [15]. This significant milestone highlights the emergence of digital-microfluidic biochip (DMFB) technology for commercial exploitation and its potential for point-of-care diagnosis [16], proteomic sample processing [17], and cell-based assays.

2.2 "Operating Systems" Research for Microfluidics and Our Vision to Expand into IoT Space

When the potential of microfluidic biochips was recognized in the late 90s, concerns were raised that higher design complexity will have to be addressed due to the need for multiple and concurrent bioassays on the chip. For example, inexpensive biochips for clinical diagnostics will integrate hematology, pathology, molecular diagnostics, cytology, microbiology, and serology onto the same platform. Motivated by this vision, efforts emerged in the research community to identify synergies between biochips and architectures for computing systems, and a number of automated design and optimization techniques were developed [18], [19]. In the past few years, a new line of innovative thinking has emerged in this area, which is driven by the need to design cyberphysical microfluidic biochips that provide tight coupling between the microfluidic hardware platform, integrated sensors, and the control software. Such cyberphysical systems allow dynamic adaptation for more flexible biochemistry-on-chip and error recovery on the fly [20].

Past research on software for digital microfluidics focused on scheduling of fluidic operations, resource binding, droplet routing, etc. [21], [22]. Algorithms for on-chip sample preparation [23], cross-contamination avoidance [24], designs for protein crystallization [25], and optimization methods for protocols such as polymerase chain reaction (PCR) have been developed [26]. Real-life demonstrations of the interplay between hardware and software in the biochip platform have been displayed in [27], [28]. These demos highlight autonomous cyberphysical operation for error recovery; i.e., they support hardware/software co-design for lab-on-chip.

While previous methods are limited to simple droplet manipulation, there is now a need to advance cyberphysical control of microfluidic biochips to make them useful for biologists. In order to map molecular biology procedures from the bench-top to a biochip, a real-time synthesis methodology was recently introduced to efficiently run quantitative gene-expression analysis and epigenetics [29], [30]. This methodology paves the way for cyberphysical microfluidic biochips to be widely adopted in biomolecular applications, especially for genomic association studies.

Similarly, there has been a considerable body of research in design automation of flow-based microfluidic biochips that utilize membrane-based valves for flow control [31]. Solutions for control-layer routing [32], [33] and wash optimization [34], as well as manufacturing testing and fault diagnosis [35], [36] have been investigated.

Despite the rich literature in design automation of microfluidic biochips, the incorporation of these biochips in an IoT-enabled framework poses new challenges at different design levels; namely the microfluidic, middleware, and application layers. Besides biochip-level cyberphysical control, a new infrastructure for cloud-level cyberphysical adaptation is required to support the complexity and scalability of genomic association studies. Furthermore, integrative methods of biology tracking and analysis across thousands of microfluidic "nodes" become necessary for efficient coordiation. In this paper, we investigate several design challenges, at different architectural layers, and potential solutions that can make BioCyBig a reality.

2.3 Biological Pathway of Gene Expression and Omic Data

Genomic association mechanisms and the associated omic data are linked to different stages of the gene-expression pathway. In the gene-expression process (Fig. 3), a particular segment of DNA is enzymatically processed, or transcribed, into an RNA molecule. Then, a specific product of RNA, namely messenger RNA (mRNA), can be expressed and contribute to the translation process. This step leads to proteins that form the functions of our life. Notably, the DNA sequences involved in the establishment of proteins are said to be "expressed genes". On the other hand, the DNA sequences that do not elucidate a high level of expression (i.e., sequences that are not regularly transcribed or are under the influence of "epigenetic transcriptional control") are said to be "silenced/suppressed genes".

Heterogeneous omic data exist within and between stages of gene-expression pathway [37]. For example, analysis of single-nucleotide polymorphism (SNP) and copy-number variation (CNV) can be applied at the genome level. Next, DNA methylation, histone modification, and chromatin accessibility are investigated at the epigenome level. Note that genes with similar DNA sequence (genome level) may not necessarily show the same expression behaviour due to variation in chromatin accessibility (epigenome level); see Fig. 3. In fact, silencing of a certain gene through tight chromatin packaging is enforced by a protein expressed from another gene located upstream or downstream from the suppressed gene. The search for such complicated inter-

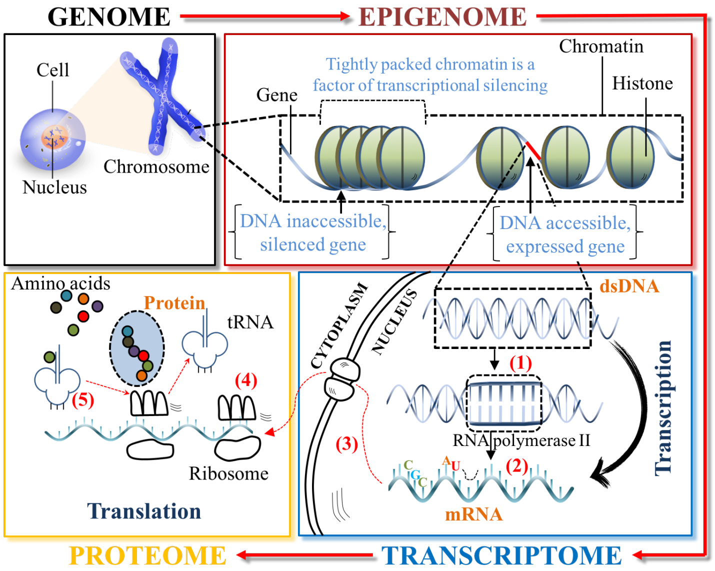

Fig. 3: The biological pathway inside a cell, and the phases of omic-data generation: genome, epigenome, transcriptome, and proteome.

action among thousands of genes is what motivates the need for BioCyBig.

At the transcriptome level, gene-expression analysis can be carried out, whereas analysis of protein expression or post-translational modification can be conducted at the proteome level. Finally, at the far end of the biological pathway, metabolite profiling in serum, plasma, etc. can be employed. To extract a certain type of omic data, dedicated bioassay protocols have been implemented using microfluidics in an isolated environment—Table 1 lists examples of such miniaturized protocols. However, only an integrative, multi-omics study of biological systems from genome, epigenome, transcriptome, proteome, and metabolome can lead to the identification of phenome; that is key to recognize serious diagnostic conditions, such as cancer or metabolic syndrome [38]. Our cyberphysical framework facilitates the coordination and management of thousdands of bioassay protocols running interactively to generate usable multi-omic data in an efficient manner. The details of the computational techniques associated with each omic data type and the integrative approaches for multi-omics are omitted due to space limitations.

Table 1: Miniaturized Protocols for Genomic-Association Studies

Pathway Omics	Biological Property	Microfluidic Protocol
Genome	SNP genotyping	Mass spectrometry [39]
Epigenome	DNA CpG methylation	Bisulfite sequencing [40]
Transcriptome	Gene expression	qPCR [41]
Proteome	Protein-DNA interactions	ChIP-seq [5]
Metabolome	Lactate release	Fluorescence-based [42]

3 Case Study: Integrative Multi-Omic Investigation of Breast Cancer

Breast cancer, like all cancer diseases, is triggered through abnormal changes in a combination of heterogeneous, yet inter-related, biological processes, including gene mutations, DNA methylation, and modifications in gene regulation and metabolism. Changes in each mechanism arise due to the activity of specific genes, which need to be identified. Combining this data leads to a genomic network that explains the multivariate association model.

Our approach not only enables progressive disease models with higher resolution over time, but it also improves our understanding of the adaptive evolutionary changes of cancer diseases [43], specially when geographically scattered microfluidic devices are involved. A special case of our progressive methodology has been proposed for studying metabolic models in biological systems [44].

Herein, we present a simplified case study that elucidates the need for an integrative multi-omic analysis for investigating breast cancer [37]. We also explain BioCyBig's role in exploring complex models of such a disease.

3.1 Multi-Omics of Breast Cancer

The goal of the study is to pinpoint the root causes of breast cancer (i.e., biomarkers). As a first step, thousands of cancerous cells must be extracted from fresh tumor tissue. Next, these cells need to be run while observing different aspects of the biological pathway (e.g., targeting genomic, epigenomic, proteomic, or metabolic associations) to construct a precise disease model using the generated multi-omic data. Note that it is difficult to obtain a large number of samples from a fresh tissue at the same site; such a limitation represents one of the bottlenecks for today's analysis techniques. Obviously, an IoT-enabled, microfluidics-driven service facilitates data integration and coordination among multiple sites.

To identify the breast-cancer model, we need to measure and integrate four types of omic data: common genetic variants (genome level), DNA methylation (epigenome level), gene expression (transcriptome level), and protein expression (proteome level). The problem objective is to construct a representative breast-cancer model based on these omics data, where gene expression is co-regulated by both DNA methylation and genetic variants. This model can be used as a disease signature to identify patients with similar tumor characteristics via clustering techniques. Thus, the model description is given below:

- **[Constraints]** Number of cancerous samples extracted from fresh tumor tissue per site.
- **[Variables]** x : Selected gene probes; y : SNPs around each gene probe per window size* (genomic data); z : CpGs around each gene probe per window size (epigenomic data); w : protein expression (proteomic data).
- **[Output]** f : Gene expression (transcriptomic data).
- **[Integrative Analysis]** Multi-staged, concatenation-based regression techniques†.

Fig. 4 shows the multi-omic analysis flow for breast cancer investigation [37]—we apply this flow to both cancer and normal cells for comparison. First, genotyping of tumor samples is performed to select gene probes and to determine the associated SNPs per each gene probe within a pre-specified window size (e.g., 1 MB window). Next, regression techniques are applied to assess the association between each expression probe and the SNPs in single and multivariate models (e.g., SNP-CpG [38]). The SNPs of probes with increasing expression activity, such as CYP1B1 gene, *may* result in high risks of carcinogenic instances.

∗. A window encompassing the gene of interest is measured in terms of megabases (MBs).

†. A valid assumptions here is that the relationship between genotype and phenotype can be modeled in a linear manner, as is the case for SNPs associated with metabolites [37].

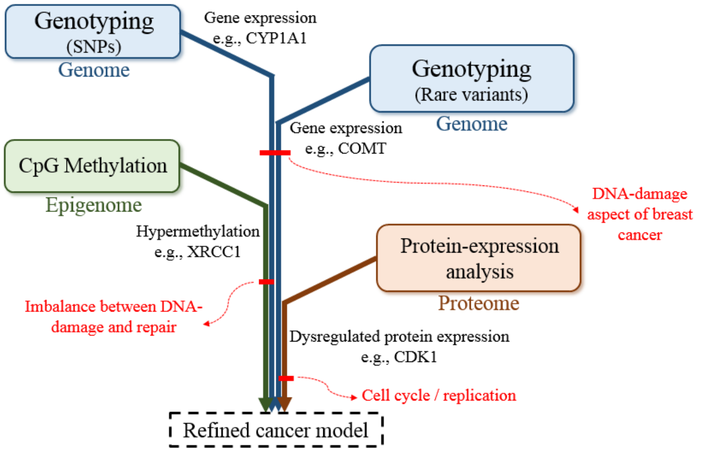

Fig. 4: Flow of integrative genomic-association analysis for breast cancer.

Likewise, genetic rare variants (or SNPs) in COMT gene can reduce the metabolism of carcinogenic product, resulting in a higher level of DNA damage. Even so, these variations may not increase the risk of cancer if the DNA-damage repair can adequately absorb carcinogenic metabolites. In other words, using variations in genetic and transcriptomic association solely as a signature for breast cancer could be misleading.

To refine the model, epigenomic and proteomic data must be integrated in the analysis flow; thus, CpG methylation data is generated and associated with gene expression. Accordingly, higher levels of methylation at XRCC1 gene and variation in the gene expression of XRCC3 result in reduced transcription levels, and the repair mechanism may no longer be able to adequately keep DNA repair at necessary levels. Even though inadequate rate of DNA-damage repair likely indicates a carcinogenic tissue, dysregulated protein expression of genes in the cell cycle pathway (e.g., CDK1) may result in a rate of cell replication that is higher than average and therefore reduces the impact of damaged cells. Hence, protein-expression analysis is equally important.

While it is evident that a study of all of the variation mentioned above is required to assess cancer development, constructing such a model requires significant quantities of samples, major effort in experimental work and interactive research, and sophisticated computation utility. These requirements can be realized using our framework.

3.2 The Use of BioCyBig

The adoption of BioCyBig as a solution for breast-cancer analysis brings the following advantages.

- BioCyBig provides unification of research goals, which enables efficient exploitation of multi-site benchtop resources (e.g., tissue samples, reagents, and workers). Such a coordination allows precise modeling of cancer, for example, through directing a research site to focus their study on specific genome loci; enabling them to increase the number of gene probes per locus and thereby the system precision. In analogy with electronic systems, this is similar to increasing the number of representation bits of an analog signal during analog-to-digital conversion.
- The big-data infrastructure can be seamlessly exploited for high-dimensional machine learning and data mining, giving a significant advantage for cancer researchers. For example, sophisticated Bayesian inference can be employed for assessing cancer risk or for predicting patient survivorship.
- Deploying BioCyBig as an open framework and reporting on constructive progress of multi-omic disease models will encourage researchers to contribute under the umbrella of BioCyBig.

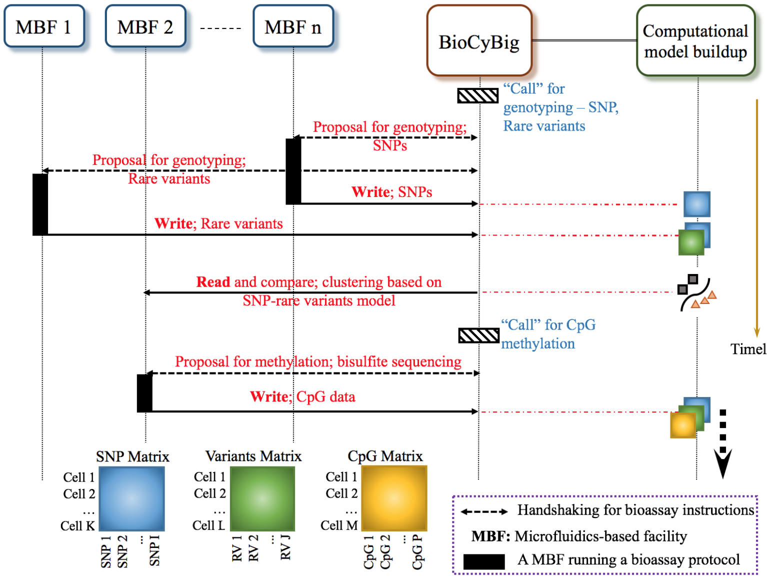

Fig. 5: Flow of integrative genomic-association analysis for breast cancer.

Fig. 5 shows the timeline of a typical scenario for the interactions between BioCyBig and breast-cancer researchers, following the logical sequence in Fig. 4. Note that a microfluidics-based facility can communicate with BioCyBig, via a handshaking mechanism, to run a bioassay protocol and augment the genomic model of a disease (i.e., "write" mode) based on a "call" from BioCyBig. Alternatively, a researcher can inquire about the current status of the model (i.e., "read" mode) for diagnosis purposes.

3.3 Working Example: CanLib

To facilitate the understanding of the BioCyBig architecture and explain the system components (Sections 5-7), we consider an example derived from the above case study of breast-cancer research. In this example, referred to as CanLib, BioCyBig is utilized to investigate the association between three types of omic data: (1) common genetic variants (via SNPs); (2) DNA methylation (via CpGs); (3) gene-expression level. The measurement of gene-expression level is performed by microfluidics-based fluorescence detection. BioCyBig collects expression data, integratively builds an association model for cancer (referred to as SNP+CpG) through a penalized regression method, assesses the significance of the model and provides cyberphysical adaptation with the aid of visual analytics. The objective is to study the expression of 21,000 gene probes, and the initial SNPs/CpGs window size for each probe is 1 Megabase. The results of the first iteration are collected from 3000 samples located at multiple sites. According to [38], it is expected that nearly 1 million SNPs and 400,000 CpGs might contribute data due to this window size. The contributing SNPs and CpGs are also referred to as transcription factors. Based on the above setting, we estimate that the size of the raw data generated in the first round to be approximately $21,000 \times 3000 \times 1,400,000 \times 4$ Bytes per entry ≈ 353 Terabytes. This value only considers static data, i.e., it does not take into

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

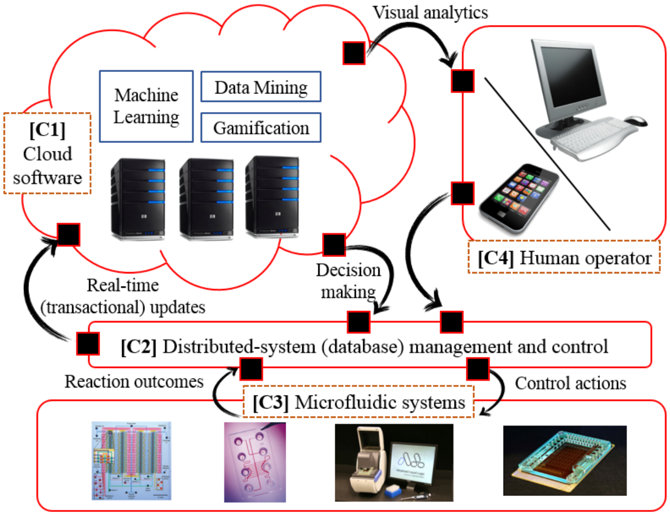

Fig. 6: Illustration of components and data flow in BioCyBig.

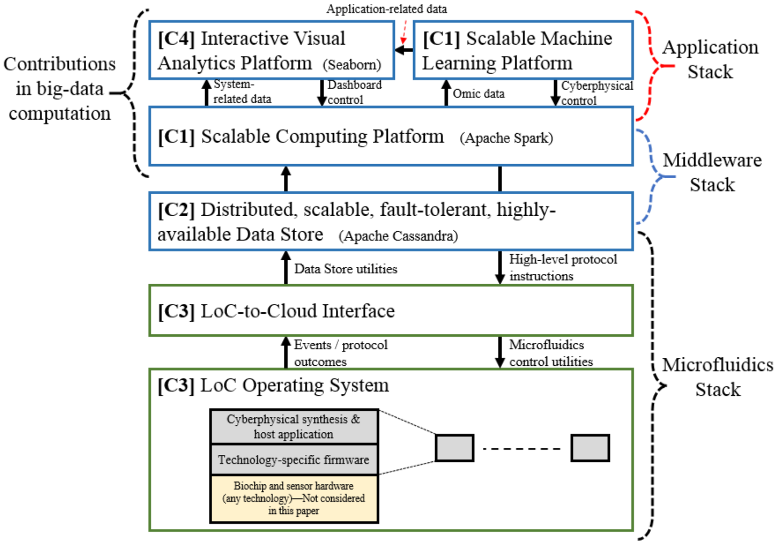
Application-related data
Contributions in big-data computation
[C4] Interactive Visual Analytics Platform (Seaborn)
[C1] Scalable Machine Learning Platform
Application Stack
System-related data
Dashboard control
Omic data
Cyberphysical control
[C1] Scalable Computing Platform (Apache Spark)
Middleware Stack
[C2] Distributed, scalable, fault-tolerant, highly-available Data Store (Apache Cassandra)
Data Store utilities
High-level protocol instructions
[C3] LoC-to-Cloud Interface
Events / protocol outcomes
Microfluidics control utilities
[C3] LoC Operating System
Microfluidics Stack
Cyberphysical synthesis & host application
Technology-specific firmware
Biochip and sensor hardware (any technology)—Not considered in this paper

Fig. 7: Software stack of BioCyBig.

account the dynamic data generated at runtime due to statistical analysis. BioCyBig is clearly motivated by the complexity of the lab procedures for this example, which considers the large number of model parameters but is nevertheless limited to only three types of omics. In practice, the association between more than three types of omic data will need to be studied.

4 The Proposed Framework: BioCyBig

In this section, we discuss the system architecture of the proposed framework.

4.1 Overall System Architecture Design

Our premise is that significant rethinking in system design is needed to leverage big-data infrastructure, CPS adaptation, and human-system interaction for distributed bioassay protocols. Fig. 6 shows a high-level view of the system components. The realization of a cloud-based microfluidic service requires the development and integration of four main components: **[C1]** cloud software infrastructure; **[C2]** distributed-system component; **[C3]** microfluidic biochip (node) component; **[C4]** human-interface component. The component **[C1]** hosts all the adaptive data-mining and machine-leaning models in the cloud. It also distributes the computational effort for constructing the multi-omics models using a cloud framework such as Apache Spark [45]. The coordination among the components **[C1]**, **[C3]** and **[C4]** are carried out at **[C2]**. The implementation of the actual microbiology protocols, either based on a feedback from **[C1]** or a human operator acting through **[C4]**, is performed through **[C3]**. Finally, the component **[C4]** incorporates human interaction, which enables a human operator to visualize the obtained results at **[C1]**, launch quick analytics procedures at runtime, and instantaneously control the distribution of biochemical assay (tasks) among **[C3]** nodes, based on technical or budgetary constraints. It is necessary to integrate these components to enable the seamless on-chip execution of complicated biochemical protocols across multiple devices, using the power of big-data analytics.

Since this is a big-data solution and intended for wide community use, the entire stack needs to be built with open-source big-data software including scalable machine learning environment such as Apache Spark, scalable, highly-available, fault-tolerant data store such as Apache Cassandra [46], and visual analytics toolings such as Standford Seaborn [47]. Fig. 7 outlines our vision for the software stack matched with the system-component schematic in Fig. 6. Blue boxes represent the components on the cloud side, whereas the green ones indicate the client components. Apache Cassandra is chosen as a distributed-data store that can directly interface with the sea of client microfluidic nodes. LoC-to-cloud interface utilities (adapters) are required to enable a microfluidic node to directly write data into Cassandra (Omic-Logging Utility). As shown in Fig. 7, Apache Spark cluster overlays on top of Cassandra to allow Spark to efficiently process complex, real-time streaming data stored in Cassandra. Apache Spark serves as our general-purpose distributed compute workhorse where many diverse multi-omics computational applications (e.g., graph data, machine learning, etc.) can be executed. To serve the visual analytics vertical application for human-system interaction, Spark will synthesize the diverse, dynamic data collected by Cassandra and write the results into Seaborn guided by a set of Python APIs that will describe visual analytics objectives. Seaborn dashboards will interactively display relevant biological and operational reports to an end user, and assist in making decisions.

4.2 Cloud Software and Human-Interface System

To improve our understanding of genomic association mechanisms, the development of novel computational tools has become an integral part of large-scale data analysis; such tools aim at converting raw data signals obtained from experimental setting to quantitative biological information. Significant effort has been devoted to addressing several challenges arising with omic data analysis [6]. For instance, due to the high dimensionality of single-cell data, enabling data visualization requires the application of specific dimensionality-reduction approaches that map the data points into a lower-dimensional space while maintaining the single-cell resolution [48]. Unsupervised clustering is widely used to group samples with similar genomic properties, which can be used to identify previously unknown subpopulations from multi-omic data. Specifying the set of genes that discriminate these subpopulations has also been studied using relevant computational tools [49]. Network modeling can provide mechanistic insights into lineage relationships and the coordination of gene activities to help in understanding the overall dynamics of the biological system. Taken together, applications of multi-omic data analysis greatly enhance the power of systematic characterization of cancer heterogeneity.

4.3 Distributed-System Architecture

An efficient coordination among multiple (heterogeneous) microfluidic devices and the cloud triggers the need for a new sensor-actor coordination model, which takes into account the specific properties of integrative multi-omics analysis. Related concepts of coordination and communication have been previously studied in systems powered with wireless sensor networks [50]. Similarly, a distributed-system design can be leveraged to collect omic data out of the microfluidic devices (sensing action). In addition, the control decisions communicated from the cloud software or the human interface are managed at this level (actuation action). At the level of large-scale genomic association analysis, the process of management and communication of control decisions is critical and it needs to be carefully studied. Several management criteria can be employed for allocating and prioritizing biochemical analysis at different microfluidic nodes. For example, a situational criterion can be used to focus biochemical analysis at certain locations in order to depict the evolution state of subpopulations, leading to an early detection of potential outbreak of highly contagious disease. Another example is the level of expertise, in which elaborate microfluidic-based biochemical analysis of liver and breast cancerous cells are performed separately at the associated research centers. Gathering information from different specialized institutions/centers will support researchers with a big picture of cancer and other diseases, and ultimately lead to a better understanding of the mechanisms behind drug resistance.

4.4 Plug-and-Play Control of Microfluidic Devices

The key technology behind the proposed cloud-based analysis system is microfluidics, which offers significant advantages for performing high-throughput screens and sensitive assays. Various microfluidic technologies (e.g., flow-based, droplet-based, and digital microfluidics) have been presented in the literature for genomic association analysis, targeting epigenome [5], transcriptome, proteome [51], etc. To seamlessly coordinate microfluidic nodes, it is necessary to develop a control methodology that is acquainted with various microfluidic technologies. A universal (canonical) control interface can be designed and customizations to specific technologies can be realized through "adapters". The design of the control interface and the adapters will include software synthesis to complement the available hardware. This standardized design of such a control interface will facilitate the plug-and-play addition of microfluidic nodes.

Logging utility tools (e.g., Omic-Logging Utility) can enable both the Lagrangian traces that record the complete fulfillment flow (e.g., which microfluidic devices worked on this biochemical experiment, when and with what outcome) and the Eulerian traces that record all state changes of a device (e.g., when this component is up, down, faulty, performing which type of operations on which experiment, and associated resource usages). This utility tool can be invoked by any biochip firmware to enable full tracing coverage in the host application.

4.5 Relationship to "Big Data" Community Goals

BioCyBig is aligned with the "Big Data" community's key goal of fostering smart and connected communities and utilizing IoT to benefit society. Our framework also offers fundamental advances into medical research by leveraging machine learning, cloud computing, and recent advances in lab-on-chip as IoT devices. It explores a completely new opportunity to rethink the principles and methods of systems engineering that are built on the foundations of real-time control, data analytics, and cyberphysical microfluidic biochips. The proposed solution investigates new engineering principles that are needed to advance personalized medicine and patient care for diseases such as cancer. It provides the means for controlling/coordinating distributed biochemistry experiments. Our choice of using open-source tools in the design is to facilitate easy adoption and technology transition. Unlike previous cyberphysical designs of microfluidic biochips, the asssessment of BioCyBig requires a combination of microfluidics-related benchmarks (e.g., technology parameters and protocols) and functional genomic data that are publicly available [52].

BioCyBig relies on three inter-related layers; namely application, middleware, and microfluidic layers. In the following three sections, we discuss design aspects and solutions of each layer. The realization of CanLib components is used as a motivating example in these sections.

5 BioCyBig Application Stack

We anticipate that multi-omic data will be collected from potentially thousands of microfluidic devices, which run biochemical experiments asynchronously. Therefore, due to the emerging complexity of the collected data, not only because of the data size but also the asynchronicity of the communicated signals, genomic-association applications (e.g., lineage inference, gene-regulatory networks, and unsupervised cell clustering) need to be populated on the cloud and accessed/delivered in the form of Software-as-a-Service (SaaS) model. To improve system productivity and engage more participants, techniques such as gamification can be used. A breakdown of the design aspects involved in this stack is given in this section.

5.1 Development of Scalable, Integrative Multi-Omic Applications on a Cloud Service

The development of novel computational tools is an integral part of genomic-association analysis; such tools nowadays come in one of two forms: (1) methods for studying the correlation between a single omic data type and gene expression; e.g., transcriptomic clustering techniques [53]; (2) integrative methods to combine different types of omics methodologies into a unified toolbox; e.g., adapted regression methods [38]. Even though such computational tools have emerged recently as an extension to single-cell benchtop work, the underlying machine learning models (highlighted below) are static and these methods rely on offline learning mechanisms. Thus there are concerns about their scalability for genome-wide association analysis and their applicability for cyberphysical adaptation. As a first step, it is necessary to analyze scalability of these techniques when multi-omic data streams (from different sources) are communicated. The evaluation metrics are computation time, computation accuracy, and system response time. Second, it is required to develop a methodology to port these offline tools into online frameworks, which interact with data from multiple sources in near-real time. Online machine-learning tools became prominent with the introduction of big data, and their role needs to be explored for genomic-association applications. Third, there is a need to design

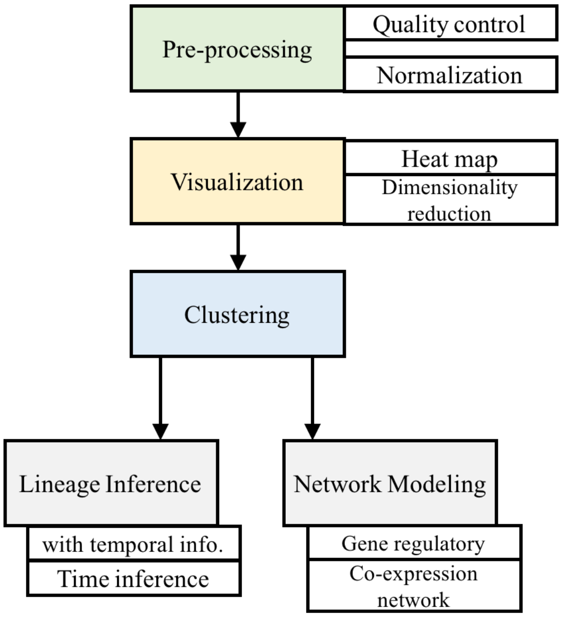

Fig. 8: Flowchart of computational single-cell analysis methods [6].

an automated delivery system that feeds data streams, received from the middleware, into the online learning models.

Fig. 8 shows a typical flowchart for single-cell genomic-association applications that need to be imported to the cloud using Apache Spark. Preprocessing and quantification are the first steps in any large-scale data analysis. The purpose of these steps is to convert raw data to quantitative biological information. In addition, significant effort is paid to the estimation and removal of systematic biases due to technical variability. A major issue in genomic analysis is that technical variation is always confounded with biological variation. Methods that aim at building error models [49] to account for biological biases, or constructing normalization techniques to correct the biases at an early stage [54], have been presented for the specific loci of a specific omic instance and they have been developed to run offline. Our solution aims at looking into algorithmic techniques to scale these methods and to incorporate cyberphysical, online adaptation feature into the application.

The high dimensionality of omic data provides a challenge for visual analytics. Several dimensionality-reduction approaches are available to map data points into a lower-dimensional space while maintaining single-cell resolution. Methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) [48] can be used to visualize omic data in different contexts. However, with technological breakthroughs in integrative, multi-omic association analysis, new distributed high-dimensional techniques are needed to keep pace with the unforeseen scaling of the cell-subpopulation features. Our proposed big-data infrastructure represents a timely advance and it provides an ideal opportunity to develop such analysis methods with unprecedented resolution. BioCyBig aims to couple algorithmic innovations with responsive big-data ecosystem leading to self-contained, dynamic computational models—this approach tends to resolve one of the biggest challenges in genomic-association studies [6].

CanLib: In addition to addressing the high dimensionality problem, CanLib must also utilize scalable machine-learning techniques that take into consideration the very small number of samples compared to the number of parameters. The most common solution to this problem is to select a subset of important explanatory variables, where the subset selection is the key of such a problem. According to [38], penalized regression allows us to accomplish this goal in a stable and computationally efficient fashion. The following multivariate model (SNP+CpG) is used: *Gene expression*$_i = \alpha_1 SNP_1 + \alpha_2 SNP_2 + ... + \alpha_1 CpG_1 + \alpha_2 CpG_2 + ...; i = 1...m$. The symbol m represents the number of gene probes.

To apply integrative analysis, the following steps are performed recursively: (1) microfluidic nodes start to perform gene-expression analysis on specified probes where a 1 MB window of SNPs and CpGs are selected from each probe; (2) the above model is modified accordingly, and penalized regression is applied to each probe and model, thereby providing the deviance per model[‡]; (3) the significance of gene probes is tested and used to guide the next iteration of probe selection. In the subsequent iterations, the window size of SNPs and CpGs is increased and only the significant probes are re-examined to update the model. This approach expands the model horizontally.

The above discussion demonstrates that cyberphysical integration and the associated big-data infrastructure offer powerful means to study the diversity and evolution of single cancer cells, which can ultimately be applied to the clinic from an early detection stage to identifying therapeutic strategies for cancer patients.

5.2 Gamification for Improving System Productivity

Improving system productivity and promoting client participation are key requirements that can play a critical role in enhancing the precision of model learning and ultimately cancer diagnosis. Therefore, inspired by Games in Health [55], the incorporation of game-design elements and strategies into the cloud side will lead to the establishment of models that will act as a key player in decision making. The cloud software needs to include the game components that will motivate the researchers/clients to gather more omic data and daily make reasonable decisions influencing positively the scope of analysis and cancer diagnosis. Gamification creates the atmosphere where more microfluidic experts become eager to participate with their experimental outcomes. As a result, a gamification software architecture (Fig. 9) must be considered as a major building block in BioCyBig.

There are several challenges that need to be addressed as part of this design aspect, as listed below:

Reward scheme: It is required to design a game where the participating users (e.g., microbiologists) are awarded based on their experimental effort, as a form of incentives. The reward scheme may be intellectual-based such as recognition in terms of academic publications and awards, or monetary-based in terms of commercialization. Automated social/academic network involvement can be used to share participants' achievements.

Mechanism design: It is needed to design the rules of the game—that can for example be taken from the principles of microeconomics to achieve fairness and efficiency among participants [56]. Features such as Pareto efficiency, envy-freeness, and sharing incentives will be potential characteristics in this game. For example, the problem of incentivizing effort and rewarding in online systems based on user contributions has evolved from

[‡] Deviance is often used in conjunction with regression to quantify the quality of fit for a model.

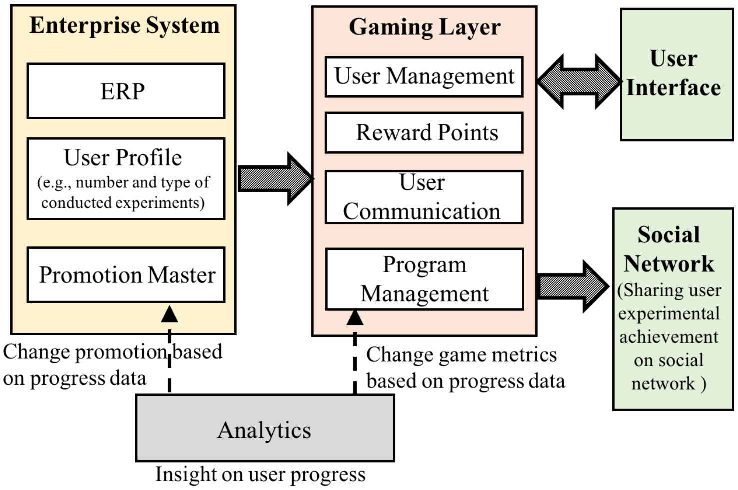
Enterprise System
ERP
User Profile
(e.g., number and type of conducted experiments)
Promotion Master
Gaming Layer
User Management
Reward Points
User Communication
Program Management
User Interface
Social Network
(Sharing user experimental achievement on social network)
Change promotion based on progress data
Change game metrics based on progress data
Analytics
Insight on user progress

Fig. 9: Proposed gamification architecture to encourage user progression and improve service productivity.

the economics literature on "tournaments" [57]. Tournaments, as a broad class of game-theoretic mechanisms, are used for diverse purposes such as choosing winners in sporting events, procuring innovations, or rewarding workers. Our innovation lies in the application of these tournaments in the cyberphysical microfluidics domain, where the players are biochemists/microbiologists utilizing microfluidics.

Game coordination and analytics: A secure and trusted software party/agent needs to be developed to coordinate/manage the game and control participants' promotion. It is necessary to investigate a game-theoretic approach for communication security (or cyber-security) to guarantee system integrity and authenticity. As a first step, potential cyber-attacks need to be explored and the actions of attackers and defenders are studied. Second, behavioral game theory can be utilized to investigate the role of certain actions taken by both parties in a set of simulated scenarios [58]. Third, reinforcement learning is used to represent a simulated attacker and a defender in cyber-security game. Such a methodology will provide us with insights about efficient techniques for ensuring cyber-security. Finally, as a part of game coordination, gamification analytics needs to be applied to our genomic analysis system.

5.3 Development of Visual Analytics and Decision Making

These tools have two main functionalities: (1) applying analytics methods to the obtained models in order to initiate automated decision making; (2) enabling users/participants to visualize and extract knowledge from the models to aid in decision making. Seaborn visualization dashboards will be adapted to be highly programmable so that they can fit different use cases. These dynamic dashboards will allow users to monitor the progression of experimental work and make protocol decisions based on their own perspective. A key challenge that needs to be addressed in this context is as follows:

Decision-making models: Decision-support systems (DSS) are used in many applications [59], and they rely on knowledge-based reasoning systems (KBS); thus KBS is adopted in BioCyBig, as shown in Fig. 10. The knowledge base will be constructed and iteratively trained (with online machine learning methods) to support decisions about protocol implementation and selection of microfluidic platforms.

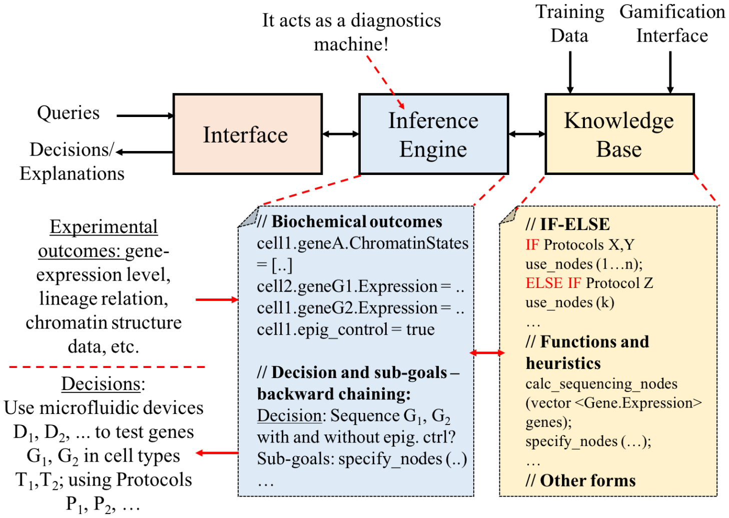
It acts as a diagnostics machine!
Training Data
Gamification Interface
Queries
Decisions/ Explanations
Interface
Inference Engine
Knowledge Base
Experimental outcomes: gene-expression level, lineage relation, chromatin structure data, etc.
Decisions: Use microfluidic devices D_1, D_2, ... to test genes G_1, G_2 in cell types T_1,T_2; using Protocols P_1, P_2, …
// Biochemical outcomes
cell1.geneA.ChromatinStates = [..]
cell2.geneG1.Expression = ..
cell1.geneG2.Expression = ..
cell1.epig_control = true
// Decision and sub-goals – backward chaining:
Decision: Sequence G_1, G_2 with and without epig. ctrl?
Sub-goals: specify_nodes (..)
…
// IF-ELSE
IF Protocols X,Y
use_nodes (1…n);
ELSE IF Protocol Z
use_nodes (k)
…
// Functions and heuristics
calc_sequencing_nodes (vector <Gene.Expression> genes);
specify_nodes (…);
…
// Other forms

Fig. 10: Components of a decision-support system.

The knowledge base can be encoded as IF-THEN, for example based on specific microfluidic technology, or it may incorporate heuristics or probabilities, for example when algorithmic partitioning of deep sequencing protocols among multiple platforms is considered. The knowledge base will be mainly stored and handled via Spark. Note that these models will run in conjunction with the designed game described above; such a pairing needs to be investigated. In other words, we are interested in investigating the influence of a decision-support mechanism on a tournament. The inference engine (or reasoning mechanism), in turn, can be developed using known concepts from artificial intelligence. It will be designed such that it receives description/analysis findings from the middleware and it may request additional information from the system user if needed. The engine will interpret the knowledge base, draw conclusions, and ultimately make decisions about protocols. In order to apply the inference engine in our setting, a backward-chaining inference methodology needs to be developed [60]—this methodology is applicable for diagnostic problems since it is a goal-directed inference, i.e., inferences about protocols or microfluidic facilities are not carried out until the system is able to reach a particular goal (e.g., execute certain protocols on specific types of cells). Fig. 10 shows an example of a cyberphysical, computer-aided decision-making scenario.

CanLib Visual Analytics: Recall that cyberphysical adaptation expands the CanLib model horizontally by re-examining the influence of transcription factors on gene probes. However, to reach a meaningful conclusion, the previous gene-expression study must be performed on different cell types, wherein cells vary based on the activity of the transcription factors. Therefore, considering multiple cell types in the study expands the CanLib model vertically. With this expansion, a user can view clustering of cells and transcriptional factors based on the level of gene expression [61]. As shown in Fig. 11(a), gene-expression analysis, obtained from multiple cell types over a specific window of transcription factors, suggests that there are two major classes of cells: C_1 and C_2. Based on Fig. 11(a), a system user/administrator can infer the following: (i) analysis over a subset of C_1 cells is sufficient to predict the overall behaviour of C_1 cells; (ii) unlike C_2 cells, C_1 cells hypothetically exhibit gene-expression activity at the north extension of transcription factors. As a result, a decision

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

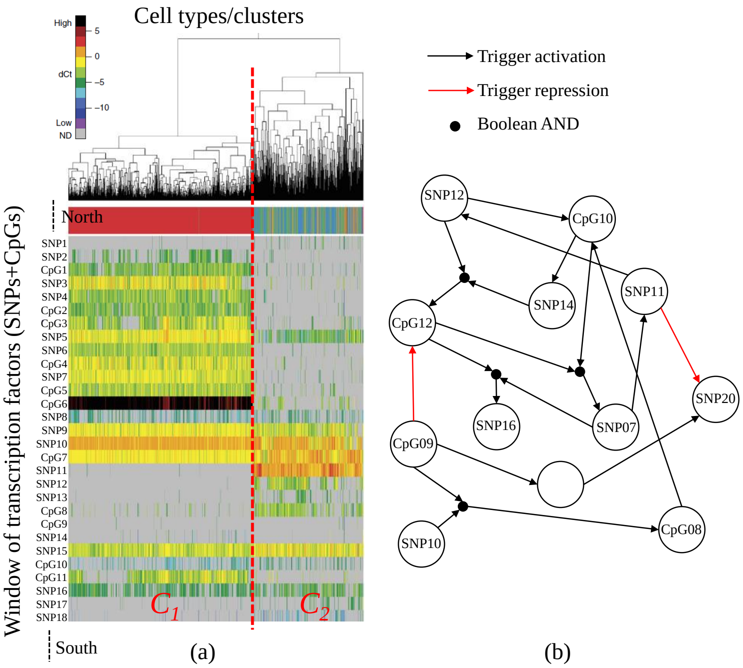

Fig. 11: Visual analytics for CanLib [61]. (a) Hierarchical clustering of cells showing the correlation between gene-expression analysis and transcription factors. (b) A synthesized Boolean network model of transcription factors illustrating the interaction between these factors (i.e., activation and repression).

can be made to extend the association analysis for a subset of C_1 by expanding the window towards the north direction.

Similarly, the synthesis of a regulatory network of transcription factors for CanLib allows researchers to infer the influence chain of each factor; see Fig. 11(b). Graph-theoretical algorithms such as finding cycles of nodes can also be used to narrow down the analysis scope.

6 Design of Microfluidics for Genomic Association Studies

Support for diverse microfluidic technologies is an important design characteristic that must be considered. Therefore, a universal control interface connected with the distributed-system software (middleware) is needed. Such an interface will enable tracing the outcomes of individual biochemical pathways and also the delivery of synthesis specifications in a technology-independent manner. Customization to specific technology, if needed, will be carried out using an "adapter" system. Fig. 12 depicts the main components of the microfluidic control and sensing interfaces.

6.1 Design of Microfluidic Control and Sensing Interface

A well-defined interface will be constructed to enable tracing and logging of experimental data. Control plans will also be conveyed through a technology-independent host controller. This level of abstraction will achieve horizontal scalability, since it can be employed at any microfluidic system. The following are some challenges that we need to consider:

Microfluidic System Model Composition: From the control perspective, our solution follows a model-verification approach at every microfluidic node to verify whether the required control decisions can be fulfilled given the advertised node model, which embeds information related to available reagents, fault-tolerance threshold, timing specifications, and others (see Fig. 12). Every participant (or researcher) will declare a unified model for the microfluidic systems they possess. Such a model will incorporate, for example, the types of tissues they work on, microfluidic technologies, and corresponding reagents, and many other features. Thus, a tool is built to turn the expert specifications into a file format that is compliant with a model-checker tool. Note that at this level, the details of the biochemical procedures are not known yet. Therefore, model verification is considered as a first step in the communication protocol between the cloud infrastructure and the underlying microfluidic systems.

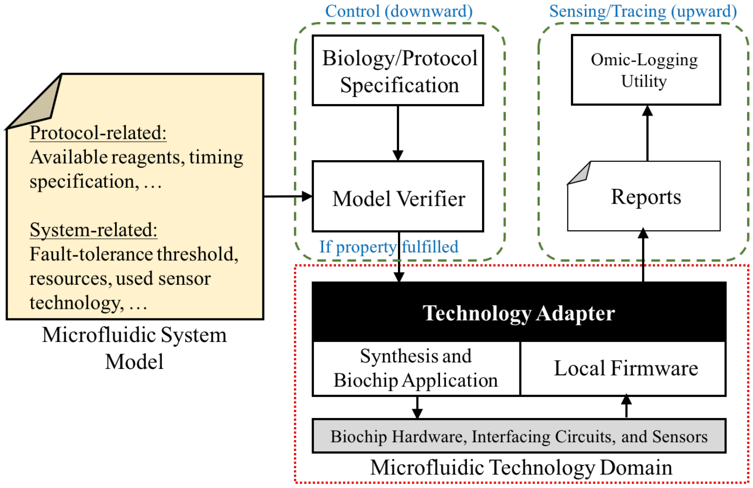
Control (downward)
Sensing/Tracing (upward)
Biology/Protocol Specification
Omic-Logging Utility
Protocol-related:
Available reagents, timing specification, …
System-related:
Fault-tolerance threshold, resources, used sensor technology, …
Microfluidic System Model
Model Verifier
Reports
If property fulfilled
Technology Adapter
Synthesis and Biochip Application
Local Firmware
Biochip Hardware, Interfacing Circuits, and Sensors
Microfluidic Technology Domain

Fig. 12: Adaptation of microfluidic platforms (clients) for BioCyBig-compliance.

Reporting: From the sensing/tracing perspective, the solution uses a reporting engine to aggregate all the details of the experimental outcomes. This data is then used by logging tools.

6.2 Omics-Driven Biochip Synthesis and Firmware

Breakthroughs in microfluidics technologies have allowed the realization of high-throughput single-cell genomic studies. The massive amount of data generated by these platforms allows mapping of complex heterogeneous tissues and most likely uncovers previously unrecognized cell types and states. The communication established between thousands of microfluidic biochips and the cloud infrastructure will interactively make use of this data. However, for automated, high-throughput, cost-effective execution of single-cell tests (e.g., deep DNA sequencing) using microfluidic biochips, there are several design and algorithmic challenges that need to be tackled first:

Scalable Protocol-DNA Co-Modeling: Recently, a graph-theoretic modeling approach for quantitative-analysis protocols has been presented [29]. While this modeling approach is capable of capturing the characteristics of a wide class of protocols (i.e., support for multiple sample pathways and sample-dependent decision making), the introduction of single-cell analysis creates new challenges (opportunities) for protocol modeling. More specifically, since a sample droplet encapsulates an individual cell, there is need to rethink protocol modeling to combine the representation of fluid-handling operations (e.g., mixing and heating) with modeling information about a particular DNA (or protein) and its associated omic data. A portion of such a protocol-DNA co-model will be specified statically, considering all possible interactions (the worst-case), while the other portion will evolve dynamically and it will guide the biochip-level

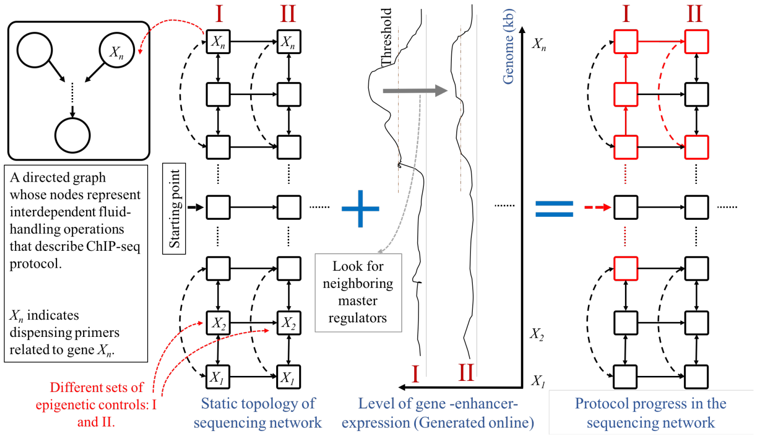
X_n
A directed graph whose nodes represent interdependent fluid-handling operations that describe ChIP-seq protocol.
X_n indicates dispensing primers related to gene X_n.
Different sets of epigenetic controls: I and II.
I
II
X_n
X_2
X_1
Starting point
Static topology of sequencing network
Threshold
Genome (kb)
Look for neighboring master regulators
Level of gene -enhancer-expression (Generated online)
Protocol progress in the sequencing network

Fig. 13: Illustration of a sequencing-network model and protocol-dynamics representation.

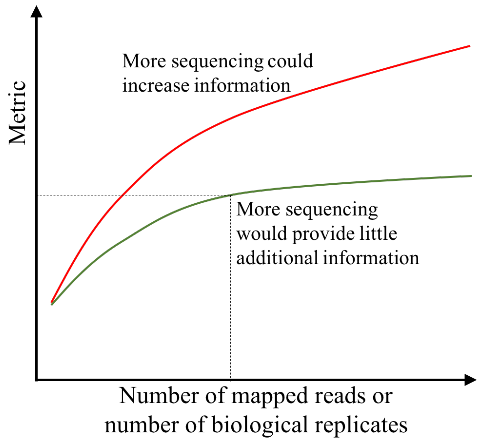

Fig. 14: Motivation for staged sequencing to predict sequencing requirements.

decision-making process to prune the search space of interactions. We call such a co-model a sequencing network, where the sequence of interactions progresses like broadcasting packets in a network. Fig. 13 shows how such a sequencing network, along with cyberphysical integration, efficiently guides biochip-level progress, on a very small scale. Execution of progression is performed through a staged synthesis framework, which is outlined below. In the course of our recent work, we realized that this model fits the requirement of many single-cell genomic-association protocols such as chromatin immunoprecipitation (ChIP) that aims to investigate DNA-protein interactions [30].

Sequencing Depth vs. Number of Droplets/Cells [Staged or Sequencing-Driven Synthesis]: Given the above protocol model as input, we propose to develop a technology-specific synthesis framework, where the progression of fluid-handling operations within the sequencing network is not known *a priori*. One of the challenging questions in next-generation sequencing associated with single-cell analysis is whether it is possible to predict the amount of sequencing that is required, both to answer a biological question and, at the same time, to prevent excessive sequencing. Fig. 14 illustrates the tradeoff between the number of cells (or replicates) required and sufficient (saturated) sequencing. Current approaches are offline; i.e., they rely on statistical analysis at the design stage. Therefore, the first problem to be handled with our framework is how CPS can enable biochip-level decision making to iteratively determine the extent of sequencing. In this case, biochip detection is applied at every step in the sequencing network. Since multiple droplets (or pathways) are involved in the protocol, the synthesis framework must consider multi-sample decision making, as presented in [29], [30]. In conclusion, our synthesis framework will combine cyberphysical control of reaction (sequencing) termination with multi-sample decision making.

The second problem arises when cell clustering and lineage network applications at the cloud layer come into play. Based on online models, a cloud-level decision making will be communicated to the biochip; the objective is to carry out deep sequencing up to a pre-determined limit. The design will aim to traverse the minimum set of fluid-handling operations to achieve the specified goal in a statistical fashion. As a first step, optimization techniques based on network-flow algorithms can be investigated to solve this problem.

Barcoding-Aware Synthesis and Firmware Design: Since thousands of cells can be involved in one experimental run in genomic association analysis, researchers have recently developed a high-throughput droplet-microfluidic approach for barcoding the RNA from thousands of individual cells for subsequent analysis by next-generation sequencing [62]. With such data, we can track heterogeneous cell sub-populations, and infer regulatory relationships between genes and pathways.

While this method provides an innovative solution to the challenging problem of cell heterogeneity and its dynamics during early differentiation, random barcoding does not allow individual cell identities (marked by gene-expression, lineage, or location) to be associated with a given barcode. In this case, sequencing efforts (to identify de novo barcoded cells) are either completely ad hoc or exhaustive, leading to extremely high completion time. Therefore, it is necessary to implement a barcoding strategy that facilitates sequencing and cell analysis at a later stage; such strategy must be incorporated in the synthesis engine and it can dynamically change the ordering of dispensed droplets accordingly. This requires appropriate reservoir/port management facility associated with microfluidic biochips. A firmware layer is also required to collect and analyze sequencing information, and to guide the next steps in the protocol.

The input to barcoding-aware synthesis will be a sequencing network and a barcoding library that represents the barcoded hydrogels. The library may contain hundreds of hydrogel barcodes, pipetted in separate reservoirs. This implies the following constraints: (1) A single port may be time-multiplexed among multiple reservoirs; (2) A multi-reservoir pressure modulator is used to control pressures at multiple reservoirs. It is obvious that such an architectural configuration may lead to significant time overhead despite being scalable and cost-effective. Our goal is to leverage this architecture, but develop a new synthesis framework (we call it barcoding-aware synthesis) that takes into account the following characteristics:

- Barcoding specifications must be fulfilled. For example, if cell differentiation is based on the gene-expression level for a gene (high, medium, and low), then we will assign a unique barcode for the cells in each category. With fine-grained gene-expression discretization and with the need to consider hundreds of genes simultaneously, barcoding becomes extremely complicated.
- Completion time must be minimized. This requires a synthesis solution for fluid handling, where multi-reservoir pres-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

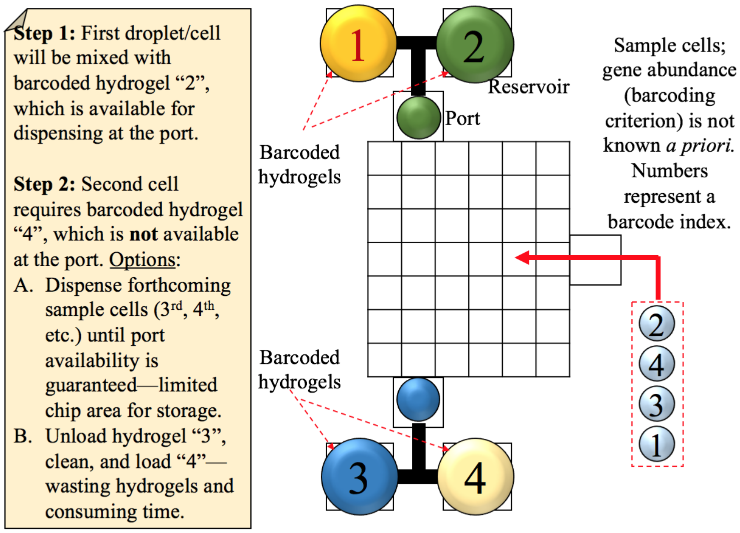
Step 1: First droplet/cell will be mixed with barcoded hydrogel "2", which is available for dispensing at the port.
Step 2: Second cell requires barcoded hydrogel "4", which is not available at the port. Options:
A. Dispense forthcoming sample cells (3rd, 4th, etc.) until port availability is guaranteed—limited chip area for storage.
B. Unload hydrogel "3", clean, and load "4"—wasting hydrogels and consuming time.
1
2
Reservoir
Port
Barcoded hydrogels
Sample cells; gene abundance (barcoding criterion) is not known a priori. Numbers represent a barcode index.
2
4
3
1
Barcoded hydrogels
3
4

Fig. 15: Motivation for barcoding-aware synthesis.

sure modulators are efficiently utilized. It is impractical to design a pressure modulator for each reservoir; the biochip may contain hundreds of reservoirs to cover a wide spectrum of barcodes. Fig. 15 illustrates the motivation for barcoding-aware synthesis.

7 Distributed-System Interfacing and Integration

The distributed-system (middleware) software will handle incoming and outgoing data streams between the microfluidic devices and the cloud, ideally in a real-time fashion. Developing the middleware stack is a big-data design problem for which similar solutions have been engineered in the past [63]. Nevertheless, this stack is important for our system and must be adopted to integrate, test, and evaluate our MaaS solution. Open-source tools, specifically Apache Spark and Cassandra, can be used for database management and analytics realization. The objective is to investigate an efficient method for data structuring with Cassandra.

7.1 Deploying Spark on Cassandra

Apache Cassandra is a masterless, NoSQL online database architecture with no single point of failure (i.e., fault tolerant). Apache Spark is a centralized scheme that designed to handle a large amount of data by simultaneously processing it at scale. For example, to develop scalable regression models for CanLib, Spark Machine Learning Library (MLlib) can be deployed and used. In BioCyBig, we tightly integrate Spark and Cassandra, which gives us the capability to use Spark to analyze the data stored at Cassandra; this data is generated online by the individual microfluidic platforms that may be geographically distributed. This integration provides horizontal scaling, fault tolerance, operational-level reporting, and analytics-friendly environment, all in one package. To achieve this integration, we will study how to extract biochemical outcomes from Cassandra and incrementally move the updates to Spark in a real-time fashion, with an online guide as a start [64].

7.2 Data Structuring and Storage in Cassandra for Real-Time Transactions

The Cassandra data model is schema-optional and column-oriented. This means that, unlike for a relational database, we do not need to model all of the columns required by the application upfront, as each row is not required to have the same set of columns. The primary language for communicating with the Cassandra database is the Cassandra Query Language (CQL). As a first step, we study how the Cassandra data model can be leveraged to store experimental reports driven by microfluidic nodes. Second, we use Cassandra APIs to write, read, and tune data replication (for consistency), via CQL. Node and cluster configurations are also an important part of our design. Our main challenge lies in the fact that biochemical reports must be structured and prepared such that real-time data writing and reading across Cassandra (via CQL) can be performed.

Integration and Evaluation: The above components need to be integrated to evaluate BioCyBig. For evaluation, publicly available information about microfluidic protocols and single-cell data can be used. Examples include a rich set of high-throughput functional genomic data a from the Bioconductor project [52]. Framework scalability can be assessed by tuning the number of microfluidic devices, labs, and users. It is required to systematically consider a range of simultaneous users and devices (e.g., in the range of tens to thousands). The responsiveness of BioCyBig can be evaluated against varying frequencies of adaptation requests.

8 Conclusion

We presented our vision for a microfluidic-driven framework, referred to as BioCyBig, that enables the interpretation of genomic sequences and how DNA mutations, expression changes, or other molecular measurements relate to disease, development, behaviour, or evolution. We illustrated the system components and their functionalities, and we explained, through a case study, how the integration of biological domain expertise, large-scale computational techniques, and a computing infrastructure can support flexible and dynamic queries and system adaption to search for patterns of genomic association over large collections of omic data.

The knowledge gained from applying molecular biology protocols to software-controlled biochips in large-scale and distributed experiments will be a big step forward towards personalized medicine. Such a cloud-infrastructure based on CPS and MaaS will advance our understanding of a variety of diseases, including cancer. The advances proposed in this paper will be applicable to a range of quantitative analysis protocols, such as gene-expression and immunological analysis.

Although this framework was originally envisioned and designed for scalable genomic and cancer research, the layered design methodology of our framework can be leveraged for other CPS areas; especially in smart city domains. For example, coupling big data analytics with cyberphysical adaptation enables management of sustainable mobility and traffic control in a smart city. In this setting, the application layer can be used for traffic data fusion, adaptive traffic-light control, and coordination of driverless transportation buses. The middleware layer, in turn, collects sensor data for traffic and weather conditions. This application and several others can be seamlessly deployed using our framework.

References

[1] T. W. House. FACT SHEET: President Obama's Precision Medicine Initiative. [Online] https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative. (Date last accessed July 22, 2016).

[2] A. Bird, "Perceptions of epigenetics," *Nature*, vol. 447, no. 7143, pp. 396–398, 2007.

[3] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: current and future challenges," *BMC systems biology*, vol. 8, no. 2, p. 1, 2014.

[4] J. F. Brothers, K. Hijazi, C. Mascaux, R. A. El-Zein, M. R. Spitz, and A. Spira, "Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-national lung screening trial era," *BMC medicine*, vol. 11, no. 1, pp. 1–15, 2013.

[5] Z. Cao, C. Chen, B. He, K. Tan, and C. Lu, "A microfluidic device for epigenomic profiling using 100 cells," *Nature Methods*, vol. 12, no. 10, pp. 959–962, 2015.

[6] A. Saadatpour, S. Lai, G. Guo, and G.-C. Yuan, "Single-cell analysis in cancer genomics," *Trends in Genetics*, vol. 31, no. 10, pp. 576–586, 2015.

[7] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: astronomical or genomical?" *PLoS Biol*, vol. 13, no. 7, pp. 1–11, 2015.

[8] T. M. SQUIRES and S. R. QUAKE, "Microfluidics: Fluid physics at the nanoliter scale," *Reviews of Modern Physics*, vol. 77, no. 3, pp. 977–1026, 2005.

[9] G. M. Whitesides, E. Ostuni, S. Takayama, X. Jiang, and D. E. Ingber, "Soft lithography in biology and biochemistry," *Annual Review of Biomedical Engineering*, vol. 3, no. 1, pp. 335–373, 2001.

[10] M. A. Unger, H.-P. Chou, T. Thorsen, A. Scherer, and S. R. Quake, "Monolithic microfabricated valves and pumps by multilayer soft lithography," *Science*, vol. 288, no. 5463, pp. 113–116, 2000.

[11] T. Thorsen, S. J. Maerkl, and S. R. Quake, "Microfluidic large-scale integration," *Science*, vol. 298, no. 5593, pp. 580–584, 2002.

[12] J. Melin and S. R. Quake, "Microfluidic large-scale integration: the evolution of design rules for biological automation," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, pp. 213–231, 2007.

[13] J. M. Perkel, "Life science technologies: microfluidicsbringing new things to life science," *Science*, vol. 322, no. 5903, pp. 975–977, 2008.

[14] R. B. Fair, "Digital microfluidics: is a true lab-on-a-chip possible?" *Microfluidics and Nanofluidics*, vol. 3, no. 3, pp. 245–281, 2007.

[15] Illumina. Illumina NeoPrep Library Prep System. [Online] http://www.illumina.com/systems/neoprep-library-system.html. (Date last accessed July 22, 2016).

[16] R. Sista, Z. Hua, P. Thwar, A. Sudarsan, V. Srinivasan, A. Eckhardt, M. Pollack, and V. Pamula, "Development of a digital microfluidic platform for point of care testing," *Lab on a Chip*, vol. 8, no. 12, pp. 2091–2104, 2008.

[17] V. N. Luk, L. K. Fiddes, V. M. Luk, E. Kumacheva, and A. R. Wheeler, "Digital microfluidic hydrogel microreactors for proteomics," *Proteomics*, vol. 12, no. 9, pp. 1310–1318, 2012.

[18] M. Ibrahim, Z. Li, and K. Chakrabarty, "Advances in design automation techniques for digital-microfluidic biochips," in *Formal Modeling and Verification of Cyber-Physical Systems*. Springer, 2015, pp. 190–223.

[19] P. Pop, W. H. Minhass, and J. Madsen, "Design methodology for flow-based microfluidic biochips," in *Microfluidic Very Large Scale Integration (VLSI)*. Springer, 2016, pp. 15–27.

[20] Y. Luo, K. Chakrabarty, and T.-Y. Ho, *Hardware/Software Co-Design and Optimization for Cyberphysical Integration in Digital Microfluidic Biochips*. Springer, 2014.

[21] Y. Zhao and K. Chakrabarty, *Design and testing of digital microfluidic biochips*. Springer Science & Business Media, 2012.

[22] C. Liao and S. Hu, "Physical-level synthesis for digital lab-on-a-chip considering variation, contamination, and defect," *IEEE Transactions on NanoBioscience*, vol. 13, no. 1, pp. 3–11, 2014.

[23] S. Roy, B. B. Bhattacharya, S. Ghoshal, and K. Chakrabarty, "Theory and analysis of generalized mixing and dilution of biochemical fluids using digital microfluidic biochips," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 11, no. 1, pp. 2:1–33, 2014.

[24] Y. Zhao and K. Chakrabarty, "Cross-contamination avoidance for droplet routing in digital microfluidic biochips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 31, no. 6, pp. 817–830, 2012.

[25] T. Xu, K. Chakrabarty, and V. K. Pamula, "Defect-tolerant design and optimization of a digital microfluidic biochip for protein crystallization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 29, no. 4, pp. 552–565, 2010.

[26] Y. Luo, B. B. Bhattacharya, T.-Y. Ho, and K. Chakrabarty, "Design and optimization of a cyberphysical digital-microfluidic biochip for the polymerase chain reaction," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 34, no. 1, pp. 29–42, 2015.

[27] K. Hu, B.-N. Hsu, A. Madison, K. Chakrabarty, and R. Fair, "Fault detection, real-time error recovery, and experimental demonstration for digital microfluidic biochips," in *Proc. IEEE/ACM Design, Automation, and Test in Europe Conference (DATE)*, 2013, pp. 559–564.

[28] K. Hu, M. Ibrahim, L. Chen, Z. Li, K. Chakrabarty, and R. Fair, "Experimental demonstration of error recovery in an integrated cyberphysical digital-microfluidic platform," in *Proc. IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2015, pp. 1–4.

[29] M. Ibrahim, K. Chakrabarty, and K. Scott, "Synthesis of cyberphysical digital-microfluidic biochips for real-time quantitative analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2016.

[30] M. Ibrahim, C. Boswell, K. Chakrabarty, K. Scott, and M. Pajic, "A real-time digital-microfluidic platform for epigenetics," in *Proc. International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES)*, 2016.

[31] P. Pop, I. E. Araci, and K. Chakrabarty, "Continuous-flow biochips: Technology, physical-design methods, and testing," *IEEE Design & Test*, vol. 32, no. 6, pp. 8–19, 2015.

[32] K. Hu, T. A. Dinh, T.-Y. Ho, and K. Chakrabarty, "Control-layer optimization for flow-based mVLSI microfluidic biochips," in *Proc. International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES)*, 2014, pp. 1–10.

[33] H. Yao, T.-Y. Ho, and Y. Cai, "PACOR: Practical control-layer routing flow with length-matching constraint for flow-based microfluidic biochips," in *Proc. IEEE/ACM Design Automation Conference (DAC)*, 2015, pp. 1–6.

[34] K. Hu, T.-Y. Ho, and K. Chakrabarty, "Wash optimization and analysis for cross-contamination removal under physical constraints in flow-based microfluidic biochips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 35, no. 4, pp. 559–572, 2016.

[35] K. Hu, F. Yu, T.-Y. Ho, and K. Chakrabarty, "Testing of flow-based microfluidic biochips: Fault modeling, test generation, and experimental demonstration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 33, no. 10, pp. 1463–1475, 2014.

[36] K. Hu, B. Bhattacharya, and K. Chakrabarty, "Fault diagnosis for leakage and blockage defects in flow-based microfluidic biochips," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 35, no. 7, pp. 1179–1191, 2016.

[37] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[38] S. Pineda, F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen, "Integration analysis of three omics data using penalized regression methods: An application to bladder cancer," *PLoS Genetics*, vol. 11, no. 12, pp. 1–22, 2015.

[39] J. Zhu, C. Qiu, M. Palla, T. Nguyen, J. J. Russo, J. Ju, and Q. Lin, "A microfluidic device for multiplex single-nucleotide polymorphism genotyping," *RSC advances*, vol. 4, no. 9, pp. 4269–4277, 2014.

[40] M. Kantlehner, R. Kirchner, P. Hartmann, J. W. Ellwart, M. Alunni-Fabbroni, and A. Schumacher, "A high-throughput DNA methylation analysis of a single cell," *Nucleic Acids Research*, vol. 39, no. 7, pp. 1–9, 2011.

[41] A. Rival, D. Jary, C. Delattre, Y. Fouillet, G. Castellan, A. Bellemin-Comte, and X. Gidrol, "An EWOD-based microfluidic chip for single-cell isolation, mRNA purification and subsequent multiplex qPCR," *Lab on a Chip*, vol. 14, no. 19, pp. 3739–3749, 2014.

[42] A. Mongersun, I. Smeenk, G. Pratx, P. Asuri, and P. Abbyad, "Droplet microfluidic platform for the determination of single-cell lactate release," *Analytical chemistry*, vol. 88, no. 6, pp. 3257–3263, 2016.

[43] M. Greaves, "Evolutionary determinants of cancer," *Cancer discovery*, vol. 5, no. 8, pp. 806–820, 2015.

[44] M. Fondi and P. Liò, "Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology," *Microbiological research*, vol. 171, pp. 52–64, 2015.

[45] T. A. S. Foundation. Apache spark. [Online] http://spark.apache.org/. (Date last accessed July 22, 2016).

[46] T. A. S. Foundation. Cassandra spark. [Online] http://cassandra.apache.org/. (Date last accessed July 22, 2016).

[47] M. Waskom. Seaborn. [Online] https://stanford.edu/∼mwaskom/software/seaborn/. (Date last accessed July 22, 2016).

[48] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[49] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods*, vol. 11, no. 7, pp. 740–742, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2643683, IEEE Transactions on Big Data

[50] L. G. Rios *et al.*, "Big data infrastructure for analyzing data generated by wireless sensor networks," in *Proc. IEEE International Congress on Big Data*, 2014, pp. 816–823.

[51] A. R. Wu, J. B. Hiatt, R. Lu, J. L. Attema, N. A. Lobo, I. L. Weissman, M. F. Clarke, and S. R. Quake, "Automated microfluidic chromatin immunoprecipitation from 2,000 cells," *Lab on a Chip*, vol. 9, no. 10, pp. 1365–1370, 2009.

[52] Bioconductor. Open source software for bioinformatics. [Online] https://www.bioconductor.org/. (Date last accessed July 22, 2016).

[53] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, 2015.

[54] K. J. Livak, Q. F. Wills, A. J. Tipping, K. Datta, R. Mittal, A. J. Goldson, D. W. Sexton, and C. C. Holmes, "Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells," *Methods*, vol. 59, no. 1, pp. 71–79, 2013.

[55] H. I. News. Games for health. [Online] http://www.healthcareitnews.com/directory/games-health. (Date last accessed July 22, 2016).

[56] M. Rabin, "Incorporating fairness into game theory and economics," *The American Economic Review*, pp. 1281–1302, 1993.

[57] D. Easley and A. Ghosh, "Incentives, gamification, and game theory: an economic approach to badge design," in *Proc. ACM Conference on Electronic Commerce (EC)*, 2013, pp. 359–376.

[58] C. F. Camerer, T.-H. Ho, and J. K. Chong, "Behavioural game theory: Thinking, learning and teaching," in *Advances in Understanding Strategic Behaviour*. Springer, 2004, pp. 120–180.

[59] D. J. Power, "Decision support systems: a historical overview," in *Handbook on Decision Support Systems 1*. Springer Berlin Heidelberg, 2008, pp. 121–140.

[60] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical informatics*. Springer, 2014, pp. 643–674.

[61] V. Moignard *et al.*, "Decoding the regulatory network of early blood development from single-cell gene expression measurements," *Nature biotechnology*, vol. 33, no. 3, pp. 269–276, 2015.

[62] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[63] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'big data', hadoop and cloud computing in genomics," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[64] DataStax. Getting started with apache spark and cassandra. [Online] https://academy.datastax.com/resources/getting-started-apache-spark-and-cassandra?unit=2217. (Date last accessed July 22, 2016).

Mohamed Ibrahim (S'13) received the B.Sc. (Hons.) degree in electrical engineering from Ain Shams University, Cairo, Egypt, in 2010, and the M.Sc. degree from the same university in 2013. Currently, he is pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering at Duke University, Durham, NC, USA.

He was appointed as a research and teaching assistant by the Faculty of Engineering, Ain Shams University, since his graduation. In 2014, he joined Prof. Chakrabarty's lab to work on the design automation and test of next-generation cyberphysical digital-microfluidic biochips. His research interests also include security aspects of microfluidic devices and big-data analytics for microbiology applications. He has contributed to digital microfluidics with 15 publications. Mohamed is a student member of IEEE and ACM.

Krishnendu Chakrabarty (F'08) received the B. Tech. degree from the Indian Institute of Technology, Kharagpur, in 1990, and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1992 and 1995, respectively. He is now the William H. Younger Distinguished Professor of Engineering in the Department of Electrical and Computer Engineering and Professor of Computer Science at Duke University. He also serves as Director of Graduate Studies for Electrical and Computer Engineering. Prof. Chakrabarty is a recipient of the National Science Foundation Early Faculty (CAREER) award, the Office of Naval Research Young Investigator award, the Humboldt Research Award from the Alexander von Humboldt Foundation, Germany, the IEEE Transactions on CAD Donald O. Pederson Best Paper award (2015), and 12 best paper awards at major conferences. He is also a recipient of the IEEE Computer Society Technical Achievement Award (2015) and the Distinguished Alumnus Award from the Indian Institute of Technology, Kharagpur (2014). He is a Research Ambassador of the University of Bremen (Germany) and a Hans Fischer Senior Fellow (named after Nobel Laureate Prof. Hans Fischer) at the Institute for Advanced Study, Technical University of Munich, Germany. He has held Visiting Professor positions at University of Tokyo and the Nara Institute of Science and Technology (NAIST) in Japan, and Visiting Chair Professor positions at Tsinghua University (Beijing, China) and National Cheng Kung University (Tainan, Taiwan).

Prof. Chakrabarty's current research projects include: testing and design-for-testability of integrated circuits; digital microfluidics, biochips, and cyberphysical systems; optimization of enterprise systems and smart manufacturing. He has authored 18 books on these topics (with one translated into Chinese and two more books in press), published over 600 papers in journals and refereed conference proceedings, and given over 260 invited, keynote, and plenary talks. He has also presented 50 tutorials at major international conferences. Prof. Chakrabarty is a Fellow of ACM, a Fellow of IEEE, and a Golden Core Member of the IEEE Computer Society. He holds nine US patents, with several patents pending. He was a 2009 Invitational Fellow of the Japan Society for the Promotion of Science (JSPS). He is a recipient of the 2008 Duke University Graduate School Dean's Award for excellence in mentoring, and the 2010 Capers and Marion McDonald Award for Excellence in Mentoring and Advising, Pratt School of Engineering, Duke University. He has served as a Distinguished Visitor of the IEEE Computer Society (2005-2007, 2010-2012), and as a Distinguished Lecturer of the IEEE Circuits and Systems Society (2006-2007, 2012-2013). Currently he serves as an ACM Distinguished Speaker.

Prof. Chakrabarty served as the Editor-in-Chief of *IEEE Design Test of Computers* during 2010-2012 and *ACM Journal on Emerging Technologies in Computing Systems* during 2010-2015. Currently he serves as the Editor-in-Chief of *IEEE Transactions on VLSI Systems*. He is also an Associate Editor of *IEEE Transactions on Computers, IEEE Transactions on Biomedical Circuits and Systems, IEEE Transactions on Multiscale Computing Systems,* and *ACM Transactions on Design Automation of Electronic Systems*. In the recent past, he has served as Associate Editor of *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2001-2013), IEEE *Transactions on Circuits and Systems I* (2005-2006), and *IEEE Transactions on Circuits and Systems II* (2010-2013).

Jun Zeng Jun Zeng is currently a principal scientist and a people manager at HP Labs in Palo Alto, California leading a research line internally named software-defined additive manufacturing. Prior to HP Labs, Jun worked as RD engineer at HP's printing group specializing in thermal and piezoelectric inkjet printheads. Prior to that he was a developer and then technical manager at Coventor (Cambridge, MA) working on developing and commercializing computer-aided design (CAD) software for micro electro-mechanical systems (MEMS) industry. From 2008 to 2013, Jun was also a termed faculty with Duke University's Department of Electrical Computer Engineering. Jun's academic training includes mechanical engineering (PhD) and computer science (M.S.), both from Johns Hopkins University. His publications include a co-edited book on CAD, a co-authored book on digital printing, 50+ peer-reviewed papers, and a dozen grand patents and several dozens more patent applications pending at USPTO. Jun is an ACM member, and an IEEE senior member.