MultiCellDS: a community-developed standard for curating microenvironment-dependent multicellular data

Authors:
Samuel H. Friedman[1], Alexander R. A. Anderson [2], David M. Bortz[3], Alexander G. Fletcher [4], Hermann B. Frieboes[5], Ahmadreza Ghaffarizadeh[1], David Robert Grimes[6], Andrea Hawkins-Daarud[7], Stefan Hoehme[8], Edwin F. Juarez[1,9], Carl Kesselman[10], Roeland M.H. Merks[11,12], Shannon M. Mumenthaler[1], Paul K. Newton[13], Kerri-Ann Norton[14], Rishi Rawat[1], Russell C. Rockne[15], Daniel Ruderman[1], Jacob Scott[16], Suzanne S. Sindi[17], Jessica L. Sparks[18], Kristin Swanson[7], David B. Agus[1], Paul Macklin[1,19,*] (corresponding author)

[1] Lawrence J. Ellison Institute for Transformative Medicine, University of Southern California, Los Angeles, CA USA
[2] Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, FL USA
[3] Applied Mathematics, University of Colorado, CO USA
[4] School of Mathematics & Statistics and Bateson Centre, University of Sheffield, Sheffield United Kingdom
[5] Bioengineering, University of Louisville, Louisville, KY USA
[6] Cancer Research UK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford United Kingdom
[7] Mathematical NeuroOncology, Mayo Clinic, Phoenix, AZ USA
[8] Institute of Computer Science, University of Leipzig, Leipzig, Germany
[9] Electrical Engineering, University of Southern California, Los Angeles, CA USA
[10] Information Sciences Institute, University of Southern California, Marina del Rey, CA USA
[11] Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[12] Mathematical Institute, Leiden University, Leiden, The Netherlands
[13] Aerospace and Mechanical Engineering, University of Southern California, Los Angeles, CA USA
[14] Systems Biology Laboratory, Johns Hopkins University, Baltimore, MD USA
[15] Mathematical Oncology, City of Hope, Duarte, CA USA
[16] Translational Hematology and Oncology Research, Cleveland Clinic, Cleveland, OH USA
[17] Applied Mathematics, University of California, Merced, CA USA
[18] Chemical, Paper, and Biomedical Engineering, Miami University, Oxford, OH USA
[19] Intelligent Systems Engineering, Indiana University, Bloomington, IN USA
[*] corresponding author: **email:** macklinp@iu.edu, **www:** http://MathCancer.org

Abstract:
Exchanging and understanding scientific data and their context represents a significant barrier to advancing research, especially with respect to information siloing. Maintaining information provenance and providing data curation and quality control help overcome common concerns and barriers to the effective sharing of scientific data. To address these problems in and the unique challenges of multicellular systems, we assembled a panel composed of investigators from several disciplines to create the MultiCellular Data Standard (MultiCellDS) with a use-case driven development process. The standard includes (1) digital cell lines, which are analogous to traditional biological cell lines, to record metadata, cellular microenvironment, and cellular phenotype variables of a biological cell line, (2) digital snapshots to consistently record simulation, experimental, and clinical data for multicellular systems, and (3) collections that can logically group digital cell lines and snapshots. We have created a MultiCellular DataBase (MultiCellDB) to store digital snapshots and the 200+ digital cell lines we have generated. MultiCellDS, by having a fixed standard, enables discoverability, extensibility, maintainability, searchability, and sustainability of data, creating biological applicability and clinical utility that permits us to identify upcoming challenges to uplift biology and strategies and therapies for improving human health.

Introduction

Over the last several years, medicine and biology have become increasingly multi-institutional and collaborative[2,11-15]. By pooling their resources, research communities can accelerate scientific progress: larger research consortia can jointly recruit patients to clinical trials that might otherwise be underpowered, create datasets that are too expensive or complex for individual institutions to build, assemble multi-parameter measurements from instruments that are too expensive or specialized to host at any one institution, and integrate diverse sets of expertise[12,14,16].

Research communities further benefit by adopting centralized data repositories with standardized data elements[20]. Sharing data and insights is less efficient when communicating by direct member-to-member interactions with non-standard data, and "insight siloing" often results[30]. Shared repositories can spread the cost of data archiving and maintenance, while data standards make it possible to jointly develop data analysis tools. When shared repositories are publicly available, they provide a gold standard dataset promoting scientific reproducibility[31] and the broader community can contribute further tools and secondary analyses that bring fresh expertise and insights[14,32,33].

Recently, a consortium of 12 cancer centres in the NCI-funded Physical Sciences in Oncology Network (PSON) performed the most comprehensive characterization of two breast cell lines to date (MCF-10A: a non-tumorigenic line, and MDA-MB-231: an aggressive metastatic line)[2]. After agreeing to unified cell line sources, culturing protocols, and quality control procedures, the consortium measured the microenvironment-dependent phenotypes of the two cell lines from multiple functional points of view (e.g., cell morphology, birth rates, multi-omics, genomic stability, mechanics, therapeutic resistance). This collaboration demonstrates that multi-institutional teams can study a problem in greater depth than would be possible by individual laboratories, with better cross-lab quality control and reproducibility.

However, this effort has not yet reached its full potential. As no standard exists for a systematic, standardized data representation, the measurements were shared in a collection of figures, spreadsheets, and documents, spread amongst individual and compressed files in websites[34] and journal-hosted supplementary data[2]. This has left the data almost undiscoverable, difficult to adapt for simulation studies and other post-publication analyses, and difficult to extend by community data contributions, in contrast to centralized repositories with standardized data elements (e.g., TCGA[35]). More broadly, multicellular systems biology is in need of an extensible, orderly data format for use in cross-lab and cross-disciplinary data exchange. Otherwise, phenotypic measurements from multicellular experiments remain trapped in individual papers, and cannot readily be mined to understand the relationships between individual cell multi-omics, microenvironment-dependent cell phenotype, multicellular dynamics, and emergent disease processes that ultimately drive patient prognosis.

In this paper, we introduce MultiCellDS (MultiCellular Data Standard): a community-developed standard to functionally describe cell phenotypes with contextual information from the microenvironment. We describe the use case-driven development process, which refined the standard until it could successfully represent a library of over 200 *digital cell lines (DCLs):* a hierarchical, standardized yet extensible representation of a biological cell line's phenotype in one or more microenvironmental contexts. Additional use cases from digital pathology and multicellular simulations drove the development of *digital snapshots*: a spatial record of the microenvironment, its cells, and their phenotypes at a single time. *Collections* allow digital snapshots and digital cell lines to be logically grouped, such as by time course (e.g., a single simulation) or study (e.g., all segmented pathology slides from a patient cohort). MultiCellDS data can include metadata to assist curation, particularly provenance (data source) when aggregating measurements from many laboratories. These standardized data elements—especially when combined with a centralised repository—can unify experimental, simulation, and clinical workflows, which often share common image and data analysis tasks, serving as biological references. See **Fig. 1**.

The MultiCellDS project is a framework for community-driven curation of biological measurements, with the goal to enable more rapid collaborative science, better quality, and enhanced reproducibility. Digital Cell Lines (DCLs) provide a standardized way of reporting what is *currently* known about context-dependent cell phenotypes and their variability, so that what is *not known* can be assessed and prioritized so as to determine the state of collective knowledge, identify the biggest gaps in that knowledge, and determine what technological advances are needed to obtain those measurements to fill in those gaps. Additionally, this project

better enables researchers to investigate the impact of intra- and inter-laboratory variability in cell culturing protocols on standardized phenotype measurements, develop standardized phenotype benchmarks and reference values, improve quality control, and contribute to better reproducibility and cross-laboratory meta-analyses of experimental findings. Together with a repository to collect and share standardized, high quality data and by adopting MultiCellDS, the community has the responsibility to use, query, add, and improve the data so we can determine which cell and tissue parameters are most predictive to cell, tissue, and disease behaviour and, via a truly systematic representation of biological knowledge, rationally prioritize research directions at national and global levels, and potentially obtain the best return for limited funding resources. By collecting multiscale, multicellular data into open, public repositories makes research more reproducible, lets scientists test new biological hypotheses and therapeutic strategies, and ultimately can lay the groundwork to improve human health.

Results

A multidisciplinary, multi-institutional team of computational modellers, biologists, clinicians, engineers, and data scientists met virtually and in person to create a standard for multicellular phenotype data via use case development. Thus, our results are both a data standard and a starting library of standardized data. (See **Method** and **Supplementary Materials** for further details.) We created three iteratively refined primary data types: a *digital cell line (DCL)*, a *digital snapshot*, and a *collections* type. We summarize these data types and their development below; full details can be found in the **Supplementary Materials** and at MultiCellDS.org.

Data type: Digital cell lines

A *digital cell line (DCL)* is a consistent, hierarchical organization of quantitative phenotype data for a single biological cell line, including the microenvironmental context of the measurements (which further organizes the data) and essential metadata ("data about data" such as biological classification, curation, quality, and citation information; see below). Its phenotype data elements are organized by function (**Table 1**) to help systemize information, while allowing flexibility and extensibility through user-defined custom data elements. The standard was refined until we could successfully represent a phenotype measurements for a library of over 200 cell types, including patient-derived and "standard" cultured cancer cell lines, endothelial cells (to test the standard against mesenchymal cells), yeast (to test against non-mammalian, non-cancer lines), and bacteria (to ensure the standard could represent prokaryotic cell data).

Each DCL has two main parts: cell line metadata and one or more phenotype datasets. Cell line metadata include information about the personnel who created and maintain a DCL, where the data came from, and other information. A phenotype dataset functionally organizes currently available cell phenotypic measurements in a specific (and annotated) microenvironmental context for the measurements (e.g. phenotype measurements in hypoxic conditions). See **Table 1**.

To record the data, we use XML (eXtensible Markup Language), a hierarchical, structured ordering of data elements that can be tailored to a specific domain using XML Schemas[36,37] and related tools. Additionally, the data can be stored in databases. XML is an ideal format for collaboratively developing MultiCellDS due to its human-editable format, the availability of many pre-existing software packages, the ecosystem of related technologies (e.g. XPATH[38]), and its widespread support across simulation and data analytics software.

Recording the cell phenotype

Cellular behaviour measurements are grouped hierarchically by function in the phenotype element. The main functional groups record parameters associated with the cell cycle (including phase information), cell death, mechanics, motility, adhesion, pharmacodynamics (denoted by PKPD to include pharmacokinetic drug metadata), transport processes of chemical entities, geometrical properties, and mass (see **Table 1**). The level of detail describing the cell cycle can vary with the experimental or clinical setup (e.g., clinical pathology often uses Ki-67 positive / negative status to quantify proliferation[39], while experimental investigations often annotate cell cycle status as G_0/G_1 or $S/G_2/M$ (e.g., using FUCCI[40]). Thus, we allow multiple cell cycle representations in a phenotype dataset. PKPD describes the pharmacokinetic of a drug response (e.g., changes in cell proliferation or motility) and drug pharmacodynamics for individual drugs (at one or more dosages) or combinations of drugs. In the future, we plan to integrate the PKPD data elements with PharmML[41].

Microenvironment: context for the phenotype

A cell's microenvironment provides the context for its phenotype by indicating the external properties of a cell, like the mechanical stress and levels of chemical substrates in the immediate vicinity of the cell. Cells often behave differently in different microenvironments. For example, the cell birth rate decreases and the mean time spent in G_0/G_1 can increase for some cell lines when oxygen is reduced from normoxic (21% O_2) conditions to hypoxic (0.1% O_2) conditions[42,43]. The microenvironmental context for a phenotype measurement set is critical, not only for comparing data across experiments and cell types, but also for building computational models[6,44]. This observational approach allows us to link microenvironment and phenotype data without fully determining subcellular mechanisms (e.g., metabolic pathways). Since the data for these microenvironmental measurements can come from multiple instruments with different sampling resolutions, or even order-of-magnitude estimates, we permit multiple domains in the microenvironment data element. The microenvironment data element defines one or more variables (e.g., extracellular oxygen concentration), annotates them against existing ontologies like Chemical Entities of Biological Interest (ChEBI)[45], and then records the value of each variable. See the Supplementary Information for more details.

Phenotype Dataset: connecting phenotype, context, and other scales

Each phenotype element is combined with a microenvironment element to form a *phenotype dataset* (a tuple of elements). This allows mapping a cell line from a position in the microenvironment space (all possible combinations of microenvironmental conditions) to a position in the phenotype space (the space of all possible phenotypes). We use keywords to denote differences between the phenotype datasets (e.g. different levels of oxygenation) to make searching between phenotype datasets easier. See the Supplementary Material for further examples. In the long term, we plan to assemble DCLs for many different phenotype datasets (breadth), with each phenotype dataset having as much multi-scale, multi-omics data as possible (depth).

Measurement-level metadata

Several metadata elements can be associated with any MultiCellDS measurement as XML attributes. The attached metadata will deal with measurement variability (e.g., standard deviation), uncertainty (e.g., margin of error), units (if applicable), and type. We annotate units with the Ontology of Units of Measure (OM)[46], which allows prefixing of units and multiplication and powers of units. Measurement type allows us to indicate whether a measurement is raw (straight off an instrument), direct (directly measured by a specialised instrument), inferred (derived by fitting a mathematical model to raw or direct data), literature (extracted from published data for the same cell type and measurement conditions), estimated (estimated to order-of-magnitude based upon published data for similar cells in similar conditions), or assumed (e.g., a mathematical simplification). See the Supplementary Material for more details.

Cell Line Metadata: information about the digital cell line

These metadata record information about an overall DCL rather than about individual measurements. Each DCL has a unique identifier (using the standards developed at identifiers.org[47]) and versioning information to help reproducibility. We record information about the people involved with a DCL (Creator, Curator, Current Contact, and Last Modified By) using well-established ORCID[48] elements. When measurements are mined from pre-existing studies, we record the source publications and analysis protocols. Next, we record biological or clinical information like what species, strain, and/or organ the cell came from, pathological or clinical classification, and any other unique identifying information. See the Supplementary Material for further details.

Provenance, Reproducibility, and Aggregated Citations

DCLs can aggregate measurements from multiple sources, using interpretations and analyses from several contributors. Over time, community-driven improvements lead to increasingly complex provenance for DCLs. The review process (see Methods) determined that DCLs must track and properly annotate this history to ensure both reproducibility and correct attribution while encouraging data contributions from the community. DCLs track data origins, creators and curators, and analysis software / algorithms as metadata. This information can be used to create an aggregated citation that properly includes this information. For example:

> "We used digital cell line CELL_LINE_NAME [R_1,R_2] version ## (MultiCellDB ##), created with data and contributions from [R_3,R_4-R_n]."

Here, R_1 is the publication that originally introduced the DCL, R_2 is the publication that introduced the latest

version (if different from R_1), MultiCellDB ## is a unique identifier in the MultiCellDB repository, R_3 is the present paper (to define the MultiCellDS data elements), and R_4-R_n are publications for (1) the original data used to create the DCLs, (2) software and algorithms used to analyse the data, and (3) previous versions of the DCL between the original and current version. This form of citation emphasizes data contributors (R_4-R_n), data curators (R_1,R_2), and analytics software developers (R_4-R_n).

Digital Cell Line Numbering and Versioning

Each DCL is assigned a unique identifier to distinguish it from other DCLs, while allowing simple identification of related DCLs. See **Fig. 2** for further details. We demonstrate the interplay of the branches and versions with curatorial control in **Fig. 3**, using concepts originating in version control software systems such as git[49] that support forks, branches, and merges of projects with version tracking.

Data type: Digital Snapshots

DCLs record averaged cell phenotype measurements and microenvironmental context on a static basis. However, it is important to record spatial and dynamical information as well. Motivated by use cases arising from recording segmented pathology images and cancer single-time simulation outputs, we extended the data elements from DCLs to create *digital snapshots*: a digitized record of the spatial distribution and phenotypic state of all cells in a specified region at a single point in time, along with spatial microenvironmental measurements. Digital snapshots can be applied to experimental, clinical or computational data, and they can describe either individual cells, clusters of cells, or cell densities. The review panel iterated on these data elements until they could sufficiently describe data outputs for several classes of computational models[50], including continuum models[28,29,51,52] cellular automata[53,54], cellular Potts models[55,56], agent-based models[51,57], and vascular network models[58,59]. See recent reviews for definitions of these models[51,57]. The panel further iterated the data elements to ensure that digital snapshots could sufficiently represent segmented pathology data from clinical and experimental samples[17]. See Supplementary Material **Human Overview** for further details.

Data type: Collections

The digital snapshot uses cases also exposed the need for logically bundling related digital snapshots, such as all saved times in a single simulation experiment or a patient cohort. Thus, we created the *collections* data type, which can bundle any combination of DCLs, digital snapshots, or other collections. The groupings permit novel and unforeseen combinations to emerge as the community continues to develop ideas about how the data should be aggregated. For example, each simulation in a computational study can now be bundled as a collection of digital snapshots (a time course), along with the DCLs used in the simulation. These collections themselves could be bundled in a collection to more easily share and archive the paper's data.

Ontology definitions, software, and extensibility

While the XML schemas specify the structure of MultiCellDS, we must also annotate the meaning of the MultiCellDS elements, allowing a consistent definition of individual terms. To that end, we generated an OWL (Web Ontology Language)[60,61] ontology from the XML schemas, permitting a mapping of each element onto its ontological counterpart. (See Box 1 to define these terms.) We extensively link to other ontologies (e.g. Gene Ontology (GO)[62,63], Cell Behavior Ontology (CBO)[64], Phenotypic Quality Ontology (PATO)[65], Ontology of Physics for Biology (OPB)[66]), as these expert-defined ontologies provide excellent, time-tested definitions. Whenever possible, we adopted definitions from these pre-existing ontologies. Where terms were non-existent or poorly suited to multicellular systems biology, we defined new terms; these are defined in the OWL files. We used CodeSynthesis' XSD/e[67] and PyXB[68] to create APIs to read and write MultiCellDS from C++ and Python respectively. We expanded the software to automatically generate OWL files from our XML schemas.

We anticipate that advances in experimental and theoretical biology and clinical science will necessitate new, descriptive data elements. To encode such data in MultiCellDS without clashing with standards-compliant software, custom data can be inserted within clearly delineated <custom> data elements anywhere in the data hierarchy. This allows scientists to rapidly adapt MultiCellDS to their needs without the potential bottleneck of requiring a full standards committee review and approval of updates to the standard. The standards panel will regularly analyse the most prevalent custom data elements for inclusion in future releases of MultiCellDS. The use of these custom elements demonstrates how MultiCellDS can seamlessly adapt to the changing needs of the research community, and how draft data elements can eventually become part of the full standard.

Digital Cell Line and Snapshot Library: MultiCellular DataBase (MultiCellDB)

We created a public repository, the MultiCellular DataBase (MultiCellDB), at http://MultiCellDS.org/MultiCellDB, in order to store the MultiCellDS files. This repository provides web based access to digital cells lines. The repository is based on the DERIVA scientific asset management system[69]. MultiCellDB allows contributors to upload DCL files which are stored in their original form, as well as converted into an entity-relationship model for storage in a relational database. The MultiCellDB interface provides a faceted search capability and allows users to search for cell line data by phenotype, measurement value, microenvironment, or metadata value. DERIVA provides a fine grain access control mechanism which allows data sets to be quarantined while they are being curated prior to public release.

To help drive development of the standard, we developed a library of over 200 DCLs. While most of these DCLs are for human cancer, we worked to ensure that the standard could handle non-cancerous and non-epithelial human cells (HUVECs), non-human mammalian cells (murine lymphoma), single celled eukaryotes (*S. cerevisiae*), and prokaryotes (*S. aureus* and *E. coli*). See **Table 2** for a complete listing and an overview of how the standard evolved due to the development of the DCLs.

While iterating on the digital snapshot standard, we recorded simulation data including: previously reported 2D agent-based ductal carcinoma in situ (DCIS) simulation data[6] an agent-based simulation of a 3D tumour spheroid using Chaste[3,4], 3D continuum simulations of vascularized brain cancer[7,8] and lymphoma[28,29], and a 2D simulation of vascularized tumour growth using BioFVM[1]. We also used digital snapshots to record nuclear morphometric properties on segmented H&E breast cancer pathology images[17]. Using the same data elements for experimental, clinical, and simulation data was a motivating factor for founding the MultiCellDS project, as it allows more direct comparison of simulation and clinical / experimental validation data, as well as more direct initialization of simulations and unified analysis platforms.

Discussion

The MultiCellDS Project has grown from a single lab's motivation to an international community, with the common goal of tackling the fundamental problem of recording and sharing multicellular data. By creating a standardized multicellular data representation with appropriate metadata, the research community can proceed to build big data repositories that are truly multiscale—including measurements from the molecular, cellular, multicellular, tissue, and whole-organism/patient scales. With a growing collection of standardized data, we can develop an ecosystem of stronger, mutually compatible software tools to quantitate and analyse data, devise new theories and predictions, and ultimately yield new discoveries that drive science and medicine.

Learning from notable prior efforts[70-72] (see **Methods** as well), we focused on the core problem of representing and structuring data, rather than annotating mathematical models. To accelerate development, we adopted a "startup" development model with a small core team responsible for leading development, an invited multidisciplinary review panel to review, refine, and test the standard, and frequent engagement with the broader scientific community through public talks, conferences, and social media outreach. We employed a use case-based development cycle (choosing test problems and brainstorm solutions; discussing the prototype data elements; testing and refining until the tests succeed) in multiple rounds. In Round 1, we worked on describing microenvironment-dependent cell phenotype and cell line metadata. Round 2 focused on describing simulation and segmented pathology data, while Round 3 (ongoing) worked to describe clinical outcome metadata and collections of data. Where possible, we leveraged existing ontologies and standards where well-developed data elements existed; we developed new data elements or extended standards when necessary. See **Supplementary Material Timeline** for additional information.

MultiCellDS' three main data types, digital cell lines (DCLs), digital snapshots, and collections, respectively allow a broad characterization of cell behaviour across many microenvironmental conditions, record spatial information (individual cells or cellular populations, distribution of microenvironmental substrates) at a single time, record key metadata, and organize the former two types into logical groupings (e.g., variants of MCF-7, time series data, data replicates). By utilizing the same data, we facilitate better exchange and comparison of data between quantitative modellers, data scientists, experimentalists, and clinicians. The use case-driven development created a library of over 200 DCLs, spanning prokaryotes and eukaryotes, epithelial and stromal cells, and patient-derived and "standard" cell lines. These are stored in a public repository, MultiCellDB, at http://MultiCellDS.org/MultiCellDB. The work also generated a collection of digitized, segmented, and spatially annotated

breast pathology data (as digital snapshots) and simulation outputs (which can be used to benchmark future models). The hierarchical nature of MultiCellDS, while challenging to traditional database methods as the data are not flat (easily represented in a two-dimensional table), is critical to maintain its organization, searchability, and extensibility.

Having agreed on a data standard with an initial public data repository, we will now turn our focus to broader community involvement, to encourage data and software tool contributions as well as refinements to the standard. This is already partly accomplished through aggregated citations: contributors are credited for contributing new measurements or improved analyses, and original sources are always attributed. The momentum of the effort may also incentivize participation: as more data and more software tools become available using the standard, new developers will have incentive to also adopt the standard to make use of those data and tools. However, at this early stage, we envision that additional incentives may be needed, such as data hackathons with prizes (to encourage creation of tools), partnerships with granting agencies and journals (to encourage publication of supplementary data in standardized formats), and collaborations with instrument makers (to enable simple, direct writing of standards-compliant data). Moreover, the repositories and software need to be searchable, flexible, and easy to use, to prevent user frustration that can ultimately doom any emerging technology. The MultiCellDS community must involve user feedback and experts in human-computer interaction when developing and refining tools. The built-in support for custom data elements should ensure sufficient flexibility, and the biologic functional-driven hierarchy of data elements helps to ensure that the data are searchable.

As we conclude the technical challenges of writing and aggregating data into shared digital cell lines, we face new challenges in curation and quality control. How do we assess the quality of a new measurement[9]? When should a new measurement replace an existing measurement? Who is responsible for and granted authority to make such decisions? These open questions can be addressed by forming community standards for curation—a natural evolution of the current standards review panel. While prior efforts exist for other biological domains (e.g. the International Society for Biocuration), the multicellular community must create curatorial data requirements suited to its own domain. As this important work progresses, the community can collectively implement heuristics to help pre-screen new data, based on prior annotated knowledge. (e.g. “Cell radius for human breast cancer cells should be between 5-15 µm; raise a warning flag and any values outside that range.”) In the future, machine learning could not only flag bad measurements but identify typical causes (e.g., culturing without a key growth factor) and suggest solutions, thus improving experimental workflows.

In the future, we will continue developing the MultiCellDS data elements and metadata to sufficiently record data from emerging clinical trials and machine learning experiments in digital pathology and will extend the standard to HDF5[73] (a widely-supported compressible hierarchical data format in the physical sciences). MultiCellDS can currently describe cell phenotype broadly across many microenvironmental conditions. Future work will add depth by integrating subcellular scales (e.g., metabolomics, receptor trafficking, and other “omics”) into the phenotype datasets, leveraging established standards such as SBML at those scales. Standardizations for the dynamics of pharmacodynamics and pharmacokinetics are needed to truly characterize high-content drug screening data and predict clinical response. We will work closely with the PharmML team to develop these data elements while harmonizing data elements across PharmML, SBML, and other standards. We will also work with leading open source computational modelling projects (particularly Chaste, CompuCell3D[55], PhysiCell[74], and TiSim[75]) to build software cross-compatibility by supporting the MultiCellDS data format; indeed, this work has already begun.

The new MultiCellDS standard and repository represent a start to sharing well-structured, machine-readable multicellular data. Now that we have tackled the issues of how to consistently record and structure data spanning many sources and types, it is time to work on building larger databases of multiscale, multicellular data, as has recently been highlighted by USA’s Vice President Biden’s “Cancer Moonshot” project[76] and countless advocates for big data, open science, and reproducibility. With large repositories of consistently structured data, we can apply machine learning, mechanistic modelling, and other quantitative techniques to the big data to start answering big and significant questions: What is the state of our collective knowledge today? Where are the biggest gaps? Which cell and tissue parameters are most predictive to cell, tissue, and disease behaviour? And if those measurements are currently lacking, what technological advances are needed to obtain them?

Through a truly systematic representation of biological knowledge, we can rationally prioritize research directions at national and global levels, and potentially get the best return for limited funding resources.

MultiCellDS gets us off the ground by providing a consistent means to record multicellular data. The repository helps to collect and share standardized data. The community now has the opportunity as well as the responsibility to make use of the data: to contribute more data; to mine the data for subtle patterns that drive new hypotheses; to test new hypotheses in computational, experimental, and clinical models; and to unlock new knowledge that drives scientific progress and yields new therapies and strategies to improve human health.

Method

Originating as MultiCellXML (MultiCellular eXtensible Markup Language)[6], MultiCellDS grew and was impacted by other standardization efforts. MultiCellDS focused on satisfying unmet needs in model-independent data descriptions—rather than model representation—to facilitate interchange of data between computational modellers, experimental biologists, and clinicians. We adopted a "startup" organizational structure to promote faster development of the data standard. We employed a use case-driven development process to iteratively assess and refine the standard while ensuring that it could succeed at its primary tasks of (1) representing and sharing microenvironment-dependent cell phenotype data while preserving provenance, and (2) sharing standardized, model-independent spatial multicellular state data from many types of simulations and segmented pathology images.

Lessons from prior standardization efforts

To date, there have been successful instances of shared biological data repositories in molecular biology (e.g., TCGA[35]) and clinical medicine (e.g., some clinical trial databases[77]), where the data are relatively homogenous (many records with identical data elements) and *flat* (non-hierarchical). However, at the intermediate scales—cell phenotype and multicellular organization—efforts have been less successful. Encyclopaedias such as the Cancer Cell Line Encyclopedia (CCLE)[78] have successfully collected useful molecular information, but have generally lacked phenotypic and multicellular data; the mapping from genotype to phenotype is not always straightforward, and demonstrating how cells interact with one another requires an understanding beyond single cells in isolation. Moreover, encyclopaedias have a tendency towards recording static data, discouraging active community contributions of improved measurements and analyses.

In a key advance towards community-driven databases, BioNumbers[7] allows much more active user contributions and provides a generalized search engine. However, the data are not organized by cell type, there is no adaptable, hierarchical organization to help searches and comparisons by function, and there is no consistent reporting of microenvironmental metadata (e.g., oxygenation conditions, and growth medium used), even as these are known to impact cell phenotype[42,43,79]. Earlier data standards have similarly worked to describe cellular data (e.g., The Systems Biology Markup Language (SBML)[70,71], CellML[72]), but they have largely focused on representing individual mathematical or computational models and their corresponding parameter values, rather than annotating model-independent cell phenotype. Moreover, these well-established standards have focused primarily on subcellular and cell-scale properties; only recently has SBML begun drafting specifications to represent multicellular models with its Dynamic and Spatial Processes packages, but the complexity of representing a wide variety of mathematical models has necessitated a long-term development process (e.g. SBML's history[80]).

The development of ontologies (e.g. GO[62,63]) has helped improve recording results, but we focus on a key difference between an ontology and a data standard: ontologies are like dictionaries by defining terms, and data standards are like grammars that give a predictable structure to the terms. Both are necessary to communicate data fully and should drive the development of one another. The Cell Behavior Ontology (CBO)[64] has helped steer the development of MultiCellDS, and we have submitted improvements to the CBO.

We chose the MultiCellDS organization model and development process based upon the lessons learned from these earlier projects. We focused on representing model-independent phenotype data to avoid the inherent complexity of representing all current and possible future mathematical models. This had the additional benefit of allowing the same data elements to represent experimental, clinical, or computational data. We iteratively tested and refined the emerging standard against a series of use cases stemming from simulation, experimental, and clinical science. This made development more concrete while ensuring that the final standard

achieved its design goals. We adopted a “startup” organizational model, where a core group of “invested” participants (based at the University of Southern California) took responsibility for drafting and developing the standard. The core group recruited a multidisciplinary review panel of leading biologists, clinicians, modellers, and data scientists, each stakeholders in their own domain to refine and improve the draft standard while completing the use cases. Public presentations at scientific meetings and social media interactions were used to solicit and incorporate feedback from the broader scientific community.

Creating the MultiCellDS standard

MultiCellDS originated as a way to save agent-based simulation data in a model of DCIS of the breast[6]. Keynote talks and discussions with modellers and experimentalists at the NCI-funded Physical Sciences in Oncology Network (PS-ON) 2013 annual meeting[81] found the need to standardize data at the multicellular scale, for improved sharing of phenotype measurements, experimental data, and simulation outputs. To drive rapid progress towards a working standard, we used a two-stage strategy: in the first year of the project, we assembled a core team at USC to integrate early consensus opinions from the 2013 PS-ON annual meeting into the preceding MultiCellXML draft specification. By the end of this year, we had developed a working draft standard for DCLs and digital snapshots.

In Year 2, we expanded the development process to a select multidisciplinary panel of computational modellers, clinicians, data scientists, software developers, and biologists. Panel members were recruited from the PS-ON and the broader scientific and medical communities. Social media engagement and public presentations were used to recruit further review panellists. The full panel membership list is given in the **Supplementary Material**. The review was organized into three rounds: Round 1 (completed) focused on DCLs, and Round 2 (completed) focused on digital snapshots for simulation and experimental data. Round 3 (in progress) is focused on clinical, pathologic, and radiologic data elements. The core team and review panel determined that the data standard should be released in parts, rather than waiting for the completion of all three rounds. This was to give faster utility, help accelerate adoption, and ease the cost and effort of software implementation, by allowing parts of the standard to be implemented as they are completed.

Each review round consisted of three phases. The first phase (“brainstorming”) used one-on-one interactions with panel members and the broader public through video conferencing and seminar talks to explore and propose data elements and their hierarchical relationships. XML schema files were maintained throughout this process to fully capture the definitions and allowed relationships in MultiCellDS XML files. The second phase (“formal review”) conducted a series of “town hall” style videoconferences with the review panel, to allow frank discussion of the elements and solicit feedback. The core group introduced refinements to the XML schema files between each town hall meeting. The third phase (“test-based refinements”) tested the new data elements against specific task(s) by the review panel members, to expose bugs and omissions in the standard. Refinements were made to the schemas until the tests could be successfully completed. See the timeline in the Supplementary Materials.

The Round 1 tests used clinical, experimental, and literature data to create DCLs for mammalian cells (cancer, stromal, and other types), non-mammalian cells (yeast), and prokaryotic cells (*S. aureus* and *E. coli*), with varying levels of detail (see **Results** and **Supplementary Material**.) Round 2 tests focused on representing simulation outputs for many discrete[3,4,6,53-56,82-84] and continuum[28,51,85-87] computational models and segmented nuclei in clinical pathology data. Round 3 tests will represent more general clinical pathology data.

Use case: Building a library of digital cell lines

We have created DCLs for a wide variety of cell types, with the largest number of DCLs coming from (1) individual patients in cancer studies and (2) many cancer types from data provided by ATCC to the PS-ON[5]. We also created DCLs for Human Umbilical Vein Endothelial Cells (HUVECs), lymphoma, yeast and bacteria. We put these DCLs into an online searchable repository (MultiCellDB). We have found that the hierarchical nature of the DCLs was challenging to traditional database methods as DCL data are not flat (easily representable in a two dimensional table); this hierarchical nature is critical to maintain the organization, searchability, and extensibility of MultiCellDS. Each collection of DCLs presented unique annotation challenges, thus driving refinements in the data standard.

Systemizing data for previous PS-ON investigation

Earlier work[2] measured a wide variety of cellular properties of MCF-10A and MDA-MB-231 cells. However, obtaining numerical information from this paper presented a set of challenges, since the data are not easily accessible. Data values are presented in plots, sometimes in the text of the paper or supplementary material, sometimes in a Word file contained in a compressed file stored on an unreferenced Wiki, and sometimes in Word documents that are not available online in any form at all (except by request from individual co-authors). We also encountered inconsistent formats and missing growth parameters. While presenting data in these various formats can speed up initial data exchange between collaborators (at the level of files, tables, and plots), it does not contribute to detailed information exchange (e.g., direct comparison and combination of measurements between groups), nor facilitate quantitative communication between computational models, other software packages, and databases. When creating the MCF-10A and MDA-MB-231 DCLs, we systemized the reported data and collected missing primary information from the authors to complete the DCLs. This systematization of the data thus continues the impetus for the original study. To successfully complete this set of DCLs, we created "dimensions" of data to record additional results. Often, a parameter is represented by a single value, but will be representative of a larger collection of non-temporal data points; for example, Young's modulus measurements as the slopes of linear fits to multiple indentation measurements. We record this larger collection to retain the original knowledge and provide better confidence in future results; see Anscombe's Quartet[88].

Creating digital cell lines from experimental protocol information

The Physical Sciences in Oncology Bioresource Core Facility offers cell line information and standard operating procedures (SOPs) for 38 human cell lines derived from 7 organs, as freely-available PDF files[5]. We extracted a growth curve (as an image) from the PDF documentation for each cell line, processed the image in MATLAB to extract the cell counts and standard deviations at several time points, and performed a nonlinear least squares fit of the extracted data to exponential and logistic growth models (see the Supplementary Materials) to extract cell birth rate parameters. (The software to reproduce this work can be found at https://sourceforge.net/projects/multicellds.) We also extracted metadata from the SOPs, and combined these with the cell birth rate parameters to create 38 minimal DCLs which can later be extended by the broader community, analogously to Wikipedia "stubs"[89]. Work on this early collection of DCLs (with its variety of cancer types and cellular origins) exposed the need for detailed biological cell line metadata elements. We introduced metadata related to a cell's origin, so that we could distinguish between the different types of cancers (e.g. breast cancer vs. colon cancer). This logically extended to have other subelements, e.g. disease or species/organisms.

Two collections of patient-derived digital cancer cell lines

Patient cohorts permit studying the variability of cell phenotype, albeit over a larger range of genetic variability. The addition of clinical data allows a systematic way of relating various phenotype measurements to clinical outcomes. We refined the MultiCellDS draft standard to record phenotype measurements for two collections of patient cell lines: glioblastoma multiforme (GBM), based upon radiology, clinical outcome, and limited pathology data[10], and DCIS of the breast, based upon morphometric and other pathology-based measurements[18]. The GBM collection includes 39 patients with motility data and an additional 133 patients with partial data, but not as complete as the other patients[10]. To successfully create the GBM DCLs, we introduced (de-identified) patient- and disease-focused metadata to the standard. For the DCIS collection, we created 17 patient-derived DCLs, using an updated calibration protocol which improved upon previous methods[6]; see the Supplementary Materials. To create these DCLs, we required new metadata elements to describe the patient clinical staging information, as well as organism, organ, and tissue identifying elements.

Other digital cell lines

The lymphoma DCLs stem from two papers[28,29], and include geometrical properties, PKPD responses, and cell cycle data. To ensure that both eukaryotic and prokaryotic cells can be successfully represented, we created DCLs for *Staphylococcus aureus (S. aureus)* and *Escherichia coli* (*E. coli*) bacteria to record proliferation, size, and mechanics data[26,27]. Future MultiCellDS releases will include new data elements for bacterial biofilms, as these properties are known to impact cell proliferation, death, and response to antibiotics and other environmental stressors[90,91]. Human Umbilical Vein Endothelial Cells (HUVECs) have several different behavioural "states": as tip, stalk, and phalanx cells[22], and we denote the different phenotype datasets using keywords. The HUVEC DCLs recorded these proliferation parameters, transport processes, motility, size and motility data elements based previous work[16,21,92,93]. Data reported in nine different papers had parameters that could have potentially been used for DCLs[21,92,94-107]; however, these papers reported qualitative data (e.g., an environmental factor increases motility, but with unspecified magnitude) or relative quantitative data (e.g., a proliferation

rate may double relative to control, but with no scale reported). This makes it difficult to combine and compare data from multiple sources.

The yeast *Saccharomyces cerevisiae* is a model organism for eukaryotic cells and cell cycling. Thus yeast are an important use case for DCLs, for building predictive mathematical models as well as translating observed microscopic behaviour to macroscopic population level behaviour. Because yeast can be grown in liquid or solid media, this DCL details the solid media preparation[23]. The replicative behaviour of yeast has been quantified in many studies[24]. We note that the cell cycle of yeast is different for cells in their first division, which have an extended G_1 phase. The replicative life span of yeast is thought to depend on the number of cell divisions rather than a specific time span; the average life span is around 25 divisions[25]. Yeast have cell walls rather than cell membranes – as in human cells – and the mechanical properties of these walls have been documented[108]. Finally, yeast cells reproduce through a budding mechanism where newly born daughter cells do not inherit the replicative age of the mother. Thus, the yeast use case required us to introduce new data elements to record a generation count that increases with each division.

Supplementary material:

The MultiCellDS project website is hosted at http://MultiCellDS.org. A list of MultiCellDB portals, including the MultiCellDB reference repository, can be found at http://portals.MultiCellDS.org.

1. User-focused overview of the data standard. [https://dx.doi.org/10.6084/m9.figshare.4269254.v1].
2. List of supported cell cycle representations. [https://dx.doi.org/10.6084/m9.figshare.4269263.v1]
3. Computer-generated documentation on the full standard, based upon the XML schema. [https://dx.doi.org/10.6084/m9.figshare.4269269].
4. The XML schema that official encodes the data standard. [https://dx.doi.org/10.6084/m9.figshare.4269272.v1 and https://dx.doi.org/10.6084/m9.figshare.4269275].
5. OWL ontology. [http://MultiCellDS.org/ont/multicellds.owl].
6. A protocol to transform DCIS pathology data into patient-derived digital cell lines. [https://dx.doi.org/10.6084/m9.figshare.4269248.v1].
7. Mathematical models used in the Chaste demonstration of MultiCellDS digital snapshots [https://dx.doi.org/10.6084/m9.figshare.4272242].
8. Matlab script used to help create ATCC-based digital cell lines [https://sourceforge.net/projects/multicellds/files/Tools/ATCC_to_digital_cell_lines/]
9. Community norms for curation, versioning, and new data elements [https://dx.doi.org/10.6084/m9.figshare.4272374.v1].
10. Current MultiCellDS invited reviewers. [http://MultiCellDS.org/Team.php#review_panel]
11. MultiCellDS invited reviewer (rounds 1-3, through Nov. 2016). [https://dx.doi.org/10.6084/m9.figshare.4272197]

All MultiCellDS documentation can be found at http://MultiCellDS.org/Documentation.php.

A list of MultiCellDS-compatible software is maintained at http://software.MultiCellDS.org.

References

1 Ghaffarizadeh, A., Friedman, S. H. & Macklin, P. BioFVM: an efficient, parallelized diffusive transport solver for 3-D biological simulations. *Bioinformatics* **32**, 1256-1258, doi:10.1093/bioinformatics/btv730 (2016).

2 The Physical Sciences-Oncology Centers Network. A physical sciences network characterization of non-tumorigenic and metastatic cells. *Scientific Reports* **3**, 1449, doi:10.1038/srep01449 http://www.nature.com/articles/srep01449 - supplementary-information (2013).

3 Pitt-Francis, J. *et al.* Chaste: A test-driven approach to software development for biological modelling. *Computer Physics Communications* **180**, 2452-2471, doi:http://dx.doi.org/10.1016/j.cpc.2009.07.019 (2009).

4 Mirams, G. R. *et al.* Chaste: An Open Source C++ Library for Computational Physiology and Biology. *PLoS Comput Biol* **9**, e1002970, doi:10.1371/journal.pcbi.1002970 (2013).

5 Physical Sciences in Oncology Network. *Background information and SOPs | Physical Sciences in Oncology*, <http://physics.cancer.gov/bioresources/SOPs.aspx> (2016).

6 Macklin, P., Edgerton, M. E., Thompson, A. M. & Cristini, V. Patient-calibrated agent-based modelling of ductal carcinoma in situ (DCIS): From microscopic measurements to macroscopic predictions of clinical progression. *Journal of Theoretical Biology* **301**, 122-140, doi:http://dx.doi.org/10.1016/j.jtbi.2012.02.002 (2012).

7 Hawkins-Daarud, A., Rockne, R., Anderson, A. & Swanson, K. Modeling Tumor-Associated Edema in Gliomas during Anti-Angiogenic Therapy and Its Impact on Imageable Tumor. *Frontiers in Oncology* **3**, doi:10.3389/fonc.2013.00066 (2013).

8 Swanson, K. R. *et al.* Quantifying the Role of Angiogenesis in Malignant Progression of Gliomas: In Silico Modeling Integrates Imaging and Histology. *Cancer Research* **71**, 7366-7375, doi:10.1158/0008-5472.can-11-1399 (2011).

9 Bastian, F. B. *et al.* The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. *Database* **2015**, doi:10.1093/database/bav043 (2015).

10 Baldock, A. L. *et al.* Invasion and proliferation kinetics in enhancing gliomas predict IDH1 mutation status. *Neuro-Oncology* **16**, 779-786, doi:10.1093/neuonc/nou027 (2014).

11 Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat Meth* **13**, 310-318, doi:10.1038/nmeth.3773 http://www.nature.com/nmeth/journal/v13/n4/abs/nmeth.3773.html - supplementary-information (2016).

12 Sellers, T. A., Caporaso, N., Lapidus, S., Petersen, G. M. & Trent, J. Opportunities and Barriers in the Age of Team Science: Strategies for Success. *Cancer Causes & Control* **17**, 229-237, doi:10.1007/s10552-005-0546-5 (2006).

13 Jones, B. F., Wuchty, S. & Uzzi, B. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science* **322**, 1259-1262, doi:10.1126/science.1158357 (2008).

14 Lee, Y.-N., Walsh, J. P. & Wang, J. Creativity in scientific teams: Unpacking novelty and impact. *Research Policy* **44**, 684-697, doi:http://dx.doi.org/10.1016/j.respol.2014.10.007 (2015).

15 Lehner, T., Senthil, G. & Addington, A. M. Convergence of Advances in Genomics, Team Science, and Repositories as Drivers of Progress in Psychiatric Genomics. *Biological Psychiatry* **77**, 6-14, doi:10.1016/j.biopsych.2014.01.003.

16 Finley, S. D., Angelikopoulos, P., Koumoutsakos, P. & Popel, A. S. Pharmacokinetics of Anti-VEGF Agent Aflibercept in Cancer Predicted by Data-Driven, Molecular-Detailed Model. *CPT: Pharmacometrics & Systems Pharmacology* **4**, 641-649, doi:10.1002/psp4.12040 (2015).

17 Dong, F. *et al.* Computational Pathology to Discriminate Benign from Malignant Intraductal Proliferations of the Breast. *PLOS ONE* **9**, e114885, doi:10.1371/journal.pone.0114885 (2014).

18 Edgerton, M. E. *et al.* A novel, patient-specific mathematical pathology approach for assessment of surgical volume: Application to ductal carcinoma in situ of the breast. *Anal. Cell. Pathol.* **34**, 247-263, doi:10.3233/ACP-2011-0019 (2011).

19 Drasdo, D. *et al.* The virtual liver: state of the art and future perspectives. *Archives of Toxicology* **88**, 2071-2075, doi:10.1007/s00204-014-1384-6 (2014).

20 Standardizing data. *Nat Cell Biol* **10**, 1123-1124 (2008).

21 Lee, E., Rosca, E. V., Pandey, N. B. & Popel, A. S. Small peptides derived from somatotropin domain-containing proteins inhibit blood and lymphatic endothelial cell proliferation, migration, adhesion and tube formation. *The international journal of biochemistry & cell biology* **43**, 1812-1821 (2011).

22 De Bock, K., De Smet, F., Leite De Oliveira, R., Anthonis, K. & Carmeliet, P. *Endothelial oxygen sensors regulate tumor vessel abnormalization by instructing phalanx endothelial cells*. Vol. 87 (2009).
23 Saghbini, M., Hoekstra, D. & Gautsch, J. *Media Formulations for Various Two-Hybrid Systems*. (Humana Press, 2001).
24 Byrne, L. J. *et al.* The Number and Transmission of [<italic>PSI</italic>⁺] Prion Seeds (Propagons) in the Yeast *Saccharomyces cerevisiae*. *PLoS ONE* **4**, e4670, doi:10.1371/journal.pone.0004670 (2009).
25 Kaeberlein, M. *et al.* Regulation of Yeast Replicative Life Span by TOR and Sch9 in Response to Nutrients. *Science* **310**, 1193-1196, doi:10.1126/science.1115535 (2005).
26 Eaton, P., Fernandes, J. C., Pereira, E., Pintado, M. E. & Xavier Malcata, F. Atomic force microscopy study of the antibacterial effects of chitosans on Escherichia coli and Staphylococcus aureus. *Ultramicroscopy* **108**, 1128-1134, doi:http://dx.doi.org/10.1016/j.ultramic.2008.04.015 (2008).
27 Lindqvist, R. Estimation of Staphylococcus aureus Growth Parameters from Turbidity Data: Characterization of Strain Variation and Comparison of Methods. *Applied and Environmental Microbiology* **72**, 4862-4870, doi:10.1128/aem.00251-06 (2006).
28 Frieboes, H. B. *et al.* An Integrated Computational/Experimental Model of Lymphoma Growth. *PLoS Comput Biol* **9**, e1003008, doi:10.1371/journal.pcbi.1003008 (2013).
29 Frieboes, H. B. *et al.* Predictive Modeling of Drug Response in Non-Hodgkin?s Lymphoma. *PLoS ONE* **10**, e0129433, doi:10.1371/journal.pone.0129433 (2015).
30 Phillips, D., Watson, L. & Willis, M. Benefits of comprehensive integrated reporting: by standardizing disparate information sources, financial executive can eliminate the narrow perspectives of the elephant and the blind man parable--and" see" beyond merely information silos or reports. *Financial Executive* **27**, 26-31 (2011).
31 Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226-1227 (2011).
32 Juty, N. *et al.* BioModels: Content, Features, Functionality, and Use. *CPT: Pharmacometrics & Systems Pharmacology* **4**, 55-68, doi:10.1002/psp4.3 (2015).
33 Wolstencroft, K. *et al.* SEEK: a systems biology data and model management platform. *BMC Systems Biology* **9**, 33, doi:10.1186/s12918-015-0174-y (2015).
34 The Physical Sciences-Oncology Centers Network. *Scientific Reports 2013 Data*, <http://psocdccwiki.utk.tennessee.edu/display/PDRI/Scientific+Reports+2013+Data> (2013).
35 The Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).
36 W3C. XML Schema Part 1: Structures Second Edition. (2004). <https://www.w3.org/TR/xmlschema-1/>.
37 W3C. XML Schema Part 2: Datatypes Second Edition. (2004). <https://www.w3.org/TR/xmlschema-2/>.
38 W3C. XML Path Language (XPath) Version 1.0. (1999). <https://www.w3.org/TR/xpath/>.
39 Scholzen, T. & Gerdes, J. The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology* **182**, 311-322, doi:10.1002/(SICI)1097-4652(200003)182:3<311::AID-JCP1>3.0.CO;2-9 (2000).
40 Sakaue-Sawano, A. *et al.* Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression. *Cell* **132**, 487-498, doi:10.1016/j.cell.2007.12.033.
41 Swat, M. J. *et al.* Pharmacometrics Markup Language (PharmML): Opening New Perspectives for Model Exchange in Drug Development. *CPT: Pharmacometrics & Systems Pharmacology* **4**, 316-319, doi:10.1002/psp4.57 (2015).
42 Mumenthaler, S. M. *et al.* The Impact of Microenvironmental Heterogeneity on the Evolution of Drug Resistance in Cancer Cells. *Cancer Informatics*, 19-31, doi:10.4137/CIN.S19338 (2015).
43 Garvey, C. M. *et al.* A high-content image-based method for quantitatively studying context-dependent cell population dynamics. *Scientific Reports* **6**, 29752, doi:10.1038/srep29752 http://www.nature.com/articles/srep29752 - supplementary-information (2016).
44 Jagiella, N., Müller, B., Müller, M., Vignon-Clementel, I. E. & Drasdo, D. Inferring Growth Control Mechanisms in Growing Multi-cellular Spheroids of NSCLC Cells from Spatial-Temporal Image Data. *PLoS Comput Biol* **12**, e1004412, doi:10.1371/journal.pcbi.1004412 (2016).
45 Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* **41**, D456-D463, doi:10.1093/nar/gks1146 (2013).

46 Rijgersberg, H., Assem, M. v. & Top, J. Ontology of units of measure and related concepts. *Semant. web* **4**, 3-13 (2013).
47 Juty, N., Le Novère, N. & Laibe, C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Research* **40**, D580-D586, doi:10.1093/nar/gkr1097 (2012).
48 Haak, L. L., Fenner, M., Paglione, L., Pentz, E. & Ratner, H. ORCID: a system to uniquely identify researchers. *Learned Publishing* **25**, 259-264, doi:10.1087/20120404 (2012).
49 Chacon, S. & Straub, B. (2014).
50 Byrne, H. & Drasdo, D. Individual-based and continuum models of growing cell populations: a comparison. *Journal of Mathematical Biology* **58**, 657-687, doi:10.1007/s00285-008-0212-0 (2008).
51 Lowengrub, J. S. *et al.* Nonlinear modelling of cancer: bridging the gap between cells and tumours. *Nonlinearity* **23**, R1 (2010).
52 Preziosi, L. *Cancer modelling and simulation*. (Chapman & Hall/CRC, 2003).
53 Anderson, A. R. A., Chaplain, M. A. J., Newman, E. L., Steele, R. J. C. & Thompson, A. M. *Mathematical Modelling of Tumour Invasion and Metastasis*. Vol. 2 129-154 (2000).
54 Poleszczuk, J. & Enderling, H. A High-Performance Cellular Automaton Model of Tumor Growth with Dynamically Growing Domains. *Applied Mathematics* **Vol.05No.01**, 9, doi:10.4236/am.2014.51017 (2014).
55 Swat, M. H. *et al.* in *Methods in Cell Biology* Vol. Volume 110 (eds R. Asthagiri Anand & P. Arkin Adam) 325-366 (Academic Press, 2012).
56 Merks, R. in *Encyclopedia of Applied and Computational Mathematics* (ed Björn Engquist) 195-201 (Springer Berlin Heidelberg, 2015).
57 Deisboeck, T. S., Wang, Z., Macklin, P. & Cristini, V. Multiscale Cancer Modeling. *Annual Review of Biomedical Engineering* **13**, 127-155, doi:doi:10.1146/annurev-bioeng-071910-124729 (2011).
58 Hoehme, S. *et al.* Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. *Proceedings of the National Academy of Sciences* **107**, 10371-10376, doi:10.1073/pnas.0909374107 (2010).
59 Friebel, A. *et al.* TiQuant: software for tissue analysis, quantification and surface reconstruction. *Bioinformatics* **31**, 3234-3236, doi:10.1093/bioinformatics/btv346 (2015).
60 W3C. OWL Web Ontology Language Reference. (2004). <https://www.w3.org/TR/owl-ref/>.
61 W3C. OWL 2 Web Ontology Language Document Overview (Second Edition). (2012). <https://www.w3.org/TR/owl2-overview/>.
62 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
63 Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**, D1049-D1056, doi:10.1093/nar/gku1179 (2015).
64 Sluka, J. P. *et al.* The cell behavior ontology: describing the intrinsic biological behaviors of real and model cells seen as active agents. *Bioinformatics*, doi:10.1093/bioinformatics/btu210 (2014).
65 Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M. & Davidson, D. Using ontologies to describe mouse phenotypes. *Genome Biology* **6**, 1-10, doi:10.1186/gb-2004-6-1-r8 (2004).
66 Cook, D. L., Bookstein, F. L. & Gennari, J. H. Physical Properties of Biological Entities: An Introduction to the Ontology of Physics for Biology. *PLoS ONE* **6**, e28708, doi:10.1371/journal.pone.0028708 (2011).
67 CC, C. S. T. *XSD/e: XML for Light-Weight C++ Applications*, <http://www.codesynthesis.com/products/xsde/> (2016).
68 Bigot, P. A. *PyXB: Python XML Schema Bindings*, <http://pyxb.sourceforge.net/> (2014).
69 R. Schuler, C. K., and K. Czajkowski. in *Proceedings of the 12th IEEE International Conference on eScience* (2016).
70 Finney, A. & Hucka, M. Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions* **31**, 1472-1473, doi:10.1042/bst0311472 (2003).
71 Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-531, doi:10.1093/bioinformatics/btg015 (2003).
72 Cuellar, A. A. *et al.* An Overview of CellML 1.1, a Biological Model Description Language. *SIMULATION* **79**, 740-747, doi:10.1177/0037549703040939 (2003).
73 Group, T. H. Hierarchical Data Format, version 5. (1997-2016).

74 Ghaffarizadeh, A., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. PhysiCell: an Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems. *bioRxiv*, doi:10.1101/088773 (2016).
75 Hoehme, S. & Drasdo, D. A cell-based simulation software for multi-cellular systems. *Bioinformatics* **26**, 2641-2642, doi:10.1093/bioinformatics/btq437 (2010).
76 Kaiser, J. & Couzin-Frankel, J. Biden seeks clear course for his cancer moonshot. *Science* **351**, 325-326, doi:10.1126/science.351.6271.325 (2016).
77 Vol. 2016 (National Library of Medicine (US), Bethesda (MD), 2000-2016).
78 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-307, doi:http://www.nature.com/nature/journal/v483/n7391/abs/nature11003.html - supplementary-information (2012).
79 Masters, J. R. & Stacey, G. N. Changing medium and passaging cell lines. *Nat. Protocols* **2**, 2276-2284 (2007).
80 Systems Biology Markup Language. *SBML Events*, <http://sbml.org/Events> (2016).
81 Macklin, P. (ed Pauline Davies) (http://physics.cancer.gov/report/2013report/PaulMacklin.aspx, 2013).
82 Ghaffarizadeh, A., Friedman, S. H. & Macklin, P. Agent-based simulation of large tumors in 3-D microenvironments. *bioRxiv*, doi:10.1101/035733 (2015).
83 Poleszczuk, J., Macklin, P. & Enderling, H. *Methods in Molecular Biology* 1-12 (Humana Press, 2016).
84 Starruß, J., de Back, W., Brusch, L. & Deutsch, A. Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics* **30**, 1331-1332, doi:10.1093/bioinformatics/btt772 (2014).
85 Macklin, P. *et al.* Multiscale modelling and nonlinear simulation of vascular tumour growth. *Journal of Mathematical Biology* **58**, 765-798, doi:10.1007/s00285-008-0216-9 (2009).
86 Spill, F., Guerrero, P., Alarcon, T., Maini, P. K. & Byrne, H. M. Mesoscopic and continuum modelling of angiogenesis. *Journal of Mathematical Biology* **70**, 485-532, doi:10.1007/s00285-014-0771-1 (2015).
87 Wu, M. *et al.* The effect of interstitial pressure on tumor growth: Coupling with the blood and lymphatic vascular systems. *Journal of Theoretical Biology* **320**, 131-151, doi:http://dx.doi.org/10.1016/j.jtbi.2012.11.031 (2013).
88 Anscombe, F. J. Graphs in Statistical Analysis. *The American Statistician* **27**, 17-21, doi:10.2307/2682899 (1973).
89 Wikipedia. *Wikipedia:Stub*, <https://en.wikipedia.org/wiki/Wikipedia:Stub> (
90 Chung, P., McNamara, P. J., Campion, J. J. & Evans, M. E. Mechanism-Based Pharmacodynamic Models of Fluoroquinolone Resistance in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy* **50**, 2957-2965, doi:10.1128/aac.00736-05 (2006).
91 Høiby, N., Bjarnsholt, T., Givskov, M., Molin, S. & Ciofu, O. Antibiotic resistance of bacterial biofilms. *International Journal of Antimicrobial Agents* **35**, 322-332, doi:http://dx.doi.org/10.1016/j.ijantimicag.2009.12.011 (2010).
92 Rosello, C., Ballet, P., Planus, E. & Tracqui, P. Model driven quantification of individual and collective cell migration. *Acta biotheoretica* **52**, 343-363 (2004).
93 Juarez, E. F. *et al.* Quantifying differences in cell line population dynamics using CellPD. *BMC Systems Biology* **10**, 92, doi:10.1186/s12918-016-0337-5 (2016).
94 Calzada, M. J. *et al.* α4β1 integrin mediates selective endothelial cell responses to thrombospondins 1 and 2 in vitro and modulates angiogenesis in vivo. *Circulation research* **94**, 462-470 (2004).
95 Cazes, A. *et al.* Extracellular Matrix–Bound Angiopoietin-Like 4 Inhibits Endothelial Cell Adhesion, Migration, and Sprouting and Alters Actin Cytoskeleton. *Circulation research* **99**, 1207-1215 (2006).
96 Finley, S., Angelikopoulos, P., Koumoutsakos, P. & Popel, A. Pharmacokinetics of Anti-VEGF Agent Aflibercept in Cancer Predicted by Data-Driven, Molecular-Detailed Model. *CPT: pharmacometrics & systems pharmacology* **4**, 641-649 (2015).
97 Hellström, M. *et al.* Dll4 signalling through Notch1 regulates formation of tip cells during angiogenesis. *Nature* **445**, 776-780 (2007).
98 Ivanov, D., Philippova, M., Tkachuk, V., Erne, P. & Resink, T. Cell adhesion molecule T-cadherin regulates vascular cell adhesion, phenotype and motility. *Experimental cell research* **293**, 207-218 (2004).
99 Kim, E. *et al.* Vasculature-specific MRI reveals differential anti-angiogenic effects of a biomimetic peptide in an orthotopic breast cancer model. *Angiogenesis* **18**, 125-136 (2015).

100 Kumar, R., Yoneda, J., Bucana, C. D. & Fidler, I. J. Regulation of distinct steps of angiogenesis by different angiogenic molecules. *International journal of oncology* **12**, 749-806 (1998).
101 Norton, K.-A., Han, Z., Popel, A. S. & Pandey, N. B. Antiangiogenic cancer drug sunitinib exhibits unexpected proangiogenic effects on endothelial cells. *OncoTargets & Therapy* **7** (2014).
102 Rosca, E. V. *et al.* A Biomimetic Collagen Derived Peptide Exhibits Anti-Angiogenic Activity in Triple Negative Breast Cancer. *PloS one* **9**, e111901 (2014).
103 Rousseau, S. *et al.* Vascular endothelial growth factor (VEGF)-driven actin-based motility is mediated by VEGFR2 and requires concerted activation of stress-activated protein kinase 2 (SAPK2/p38) and geldanamycin-sensitive phosphorylation of focal adhesion kinase. *Journal of Biological Chemistry* **275**, 10661-10672 (2000).
104 Siemerink, M. J. *et al.* CD34 marks angiogenic tip cells in human vascular endothelial cell cultures. *Angiogenesis* **15**, 151-163 (2012).
105 Song, J. W. & Munn, L. L. Fluid forces control endothelial sprouting. *Proceedings of the National Academy of Sciences* **108**, 15342-15347 (2011).
106 Stamatelos, S. K., Kim, E., Pathak, A. P. & Popel, A. S. A bioimage informatics based reconstruction of breast tumor microvasculature with computational blood flow predictions. *Microvascular research* **91**, 8-21 (2014).
107 Stefanini, M. O., Wu, F. T., Mac Gabhann, F. & Popel, A. S. The presence of VEGF receptors on the luminal surface of endothelial cells affects VEGF distribution and VEGF signaling. *PLoS Comput Biol* **5**, e1000622 (2009).
108 Stenson, J. D., Hartley, P., Wang, C. & Thomas, C. R. Determining the mechanical properties of yeast cell walls. *Biotechnology Progress* **27**, 505-512, doi:10.1002/btpr.554 (2011).

FIGURES AND TABLES

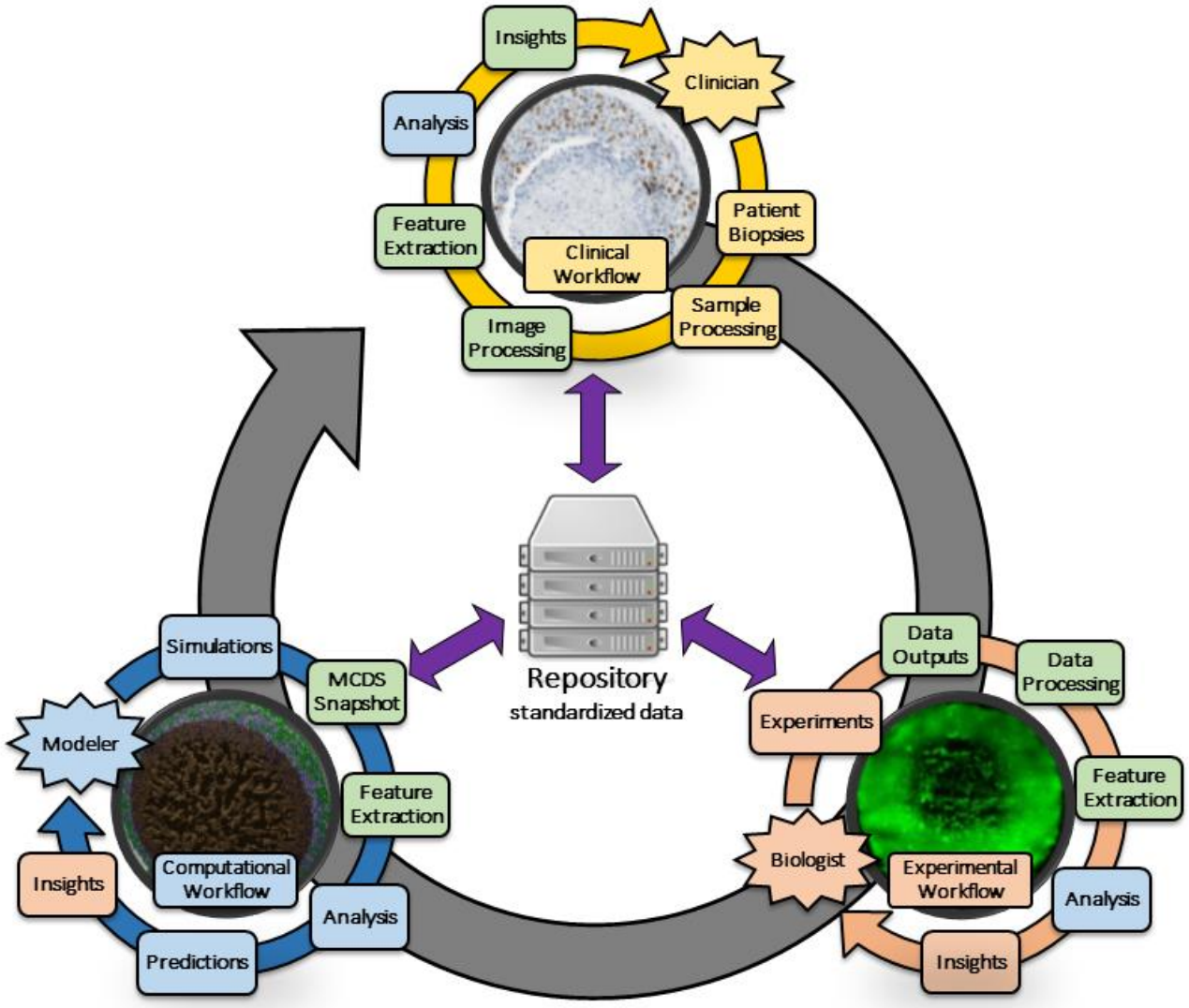

Fig. 1: Experimental, computational, and clinical workflows all share data/image analysis (green), quantitative / analytical (blue), and biological interpretive (red) tasks. Adopting standardized data representations, along with shared data repositories, allows better integration of these workflows: standards-compliant image processing, data analysis, and machine learning software can be reused across the workflows, and computational, experimental, and clinical models can be more readily compared and cross-validated.

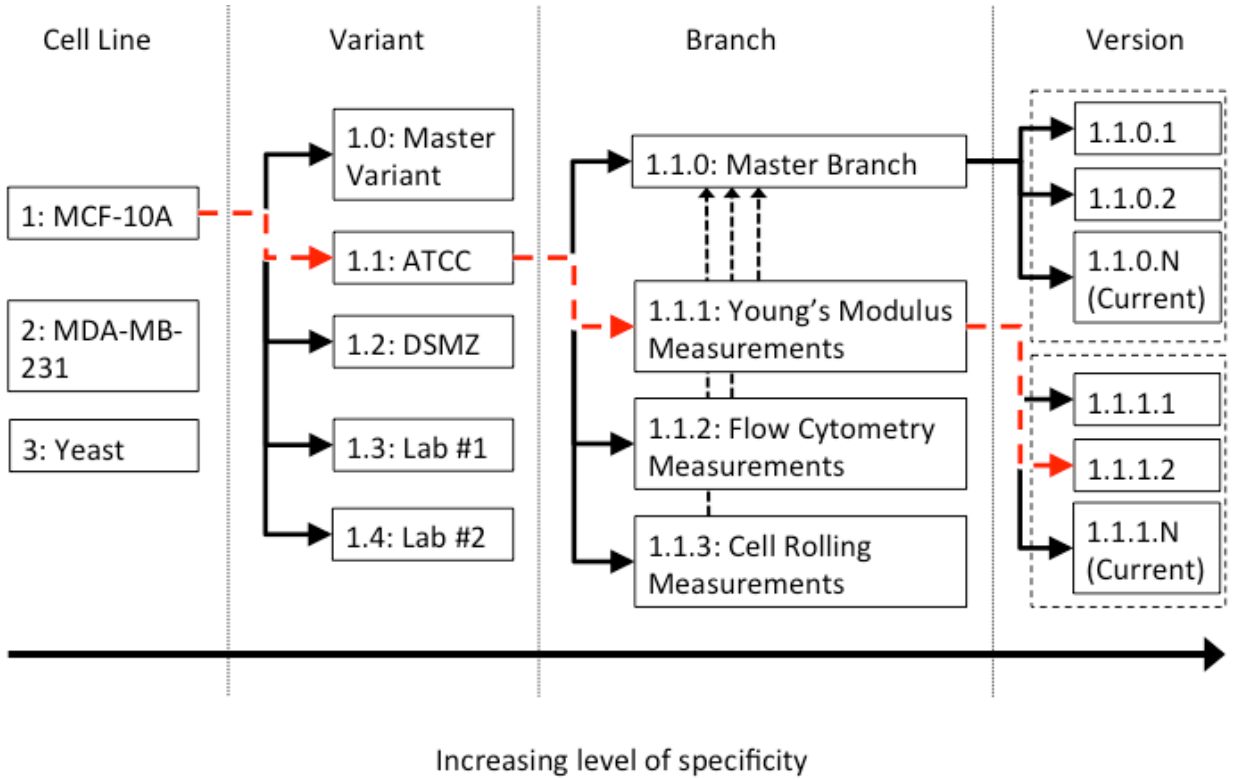

Fig 2. Systematic version-based numbering: Each number has four parts: Line.Variant.Branch.Version. A Line represents a distinct type of cells, e.g. MCF-7. Each Variant represents a (sub-)line associated with a particular organization or group that may have customized the cell line, e.g. ATCC or an investigator's group. Each list of variants is independent of the parent line. The 0th variant represents the collective behaviour aggregated over all variants. Each Branch represents a particular set of (curated or uncurated) measurements. For example, one branch could have curated Young's Modulus measurements and another branch could have uncurated cell cycle phase durations. The 0th branch is the curated version with the most data included, pending any potential conflicts in the data between various branches. The Version represents the improvement (increase in quantity or quality) of the data in the DCL, thus recording the history and evolution of the data. By having lines, variants, branches, and versions of DCLs, we can record new data without impacting any pre-existing DCLs. In this example, the red line shows DCL 1.1.1.2 for the MCF-10A line with the ATCC variant with Young's Modulus measurements and the second version of the data.

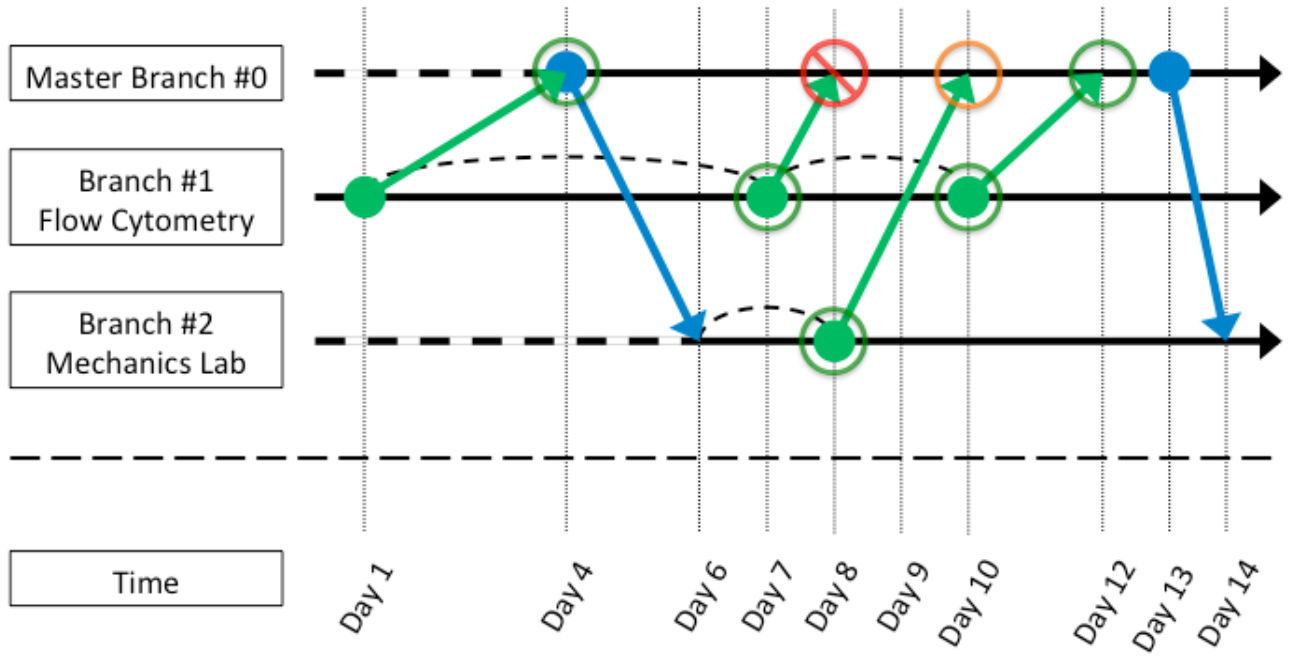

Fig 3. Interaction of branches and versions (example). Initially, a flow cytometry group contributes data in a new branch (#1) and merges it into the master branch (#0). A previously designated master branch curator accepts the merge. Later, the flow cytometry group mistakenly updates its branch (#1) with faulty data; the curator rejects the merge to the master branch. Meanwhile, a cell mechanics group forks (duplicates) the master branch to create a new mechanics branch (#2), adds their data to the new branch, and merges this data into the master branch. Because the master branch curator cannot evaluate the mechanics data and it does not conflict with any pre-existing data, the data are automatically accepted into the master branch. The flow cytometry group collects new data to revise their branch (#1), and the master branch curator approves the merge. Lastly, the mechanics group seeks the newly curated flow cytometry data and merges the master branch into their mechanics branch (#2). Each successful commit or merge results in a new version of the DCL. In terms of access, all data should be publically available, unless a private branch has been created (forking). We purposely limit describing any policies to let each community develop their own standards. See **Supplementary Material: Norms for MultiCellDS**.
Legend: Solid circles are commits, green solid upward diagonal lines are merges, blue solid downward diagonal lines are pulls, unfilled circles are conflicts and merges, and curved dashed lines are the introduction of additional data.

Main elements in a digital cell line (DCL)

Metadata:	Biological cell line details, data sources, curation and contact details, versioning, and citation information
Phenotype dataset: ○ Microenvironment ○ Phenotype	A set of measurements in a microenvironmental context ○ Details on the microenvironmental (ME) conditions ○ Phenotype measurements in this ME context

Key functional elements in a phenotype

Functional group	Typical data elements
cell cycle	Duration and variability of cell cycle phases, death rates
cell death	Duration, water loss rate, degradation rate parameters
mechanics	Young's modulus, maximum stretching deformation
adhesion	Surface adhesion receptors
motility	Maximum cell velocity, correlation with chemokines
transport	Substrate import and export rates
PKPD	Cell birth and death response to a specific drug dose
geometrical properties	Cell volume, fluid fraction
mass	Cell biomass
cell parts	Young's modulus in the nucleus

Table 1. Key functional parts of a digital cell line (top) and phenotype (bottom).

Collections of digital cell lines

Collection	Number	Description	New or Improved Elements
PSON	2	Very detailed MCF-10A and MDA-MB-231 human breast cell lines based upon PSON work.[2]	"Dimensions" of results, e.g. force vs. indentation amount
ATCC	38	cancer cell lines built from publicly-available ATCC data sheets[5]	None
GBM	39 (133)	glioblastoma multiforme (aggressive brain cancer) based on patient data[10]. 39 detailed lines, 133 additional partial lines.	Patient and Disease Elements
DCIS	17	ductal carcinoma in situ of the breast, built from previously published patient pathology data[18] and modest extensions of earlier parameter estimation protocols[6,19].	Patient and Disease Elements
endothelial	1	endothelial cell lines, built from prior literature[16,21] for HUVECs	None
yeast	1	*species* based on previously-published data[22-25] (first non-mammalian digital cell line)	Generation number
bacteria	2	*species* based on previously-published data[26,27] (first non-eukaryotic digital cell lines)	None
lymphoma	2	murine lymphoma lines based upon previously published experimental/simulation studies[28,29]	None

Table 2. Summary of digital cell lines in the library.

Collections of digital snapshots

Collection	Description	New or Improved Elements
BioFVM	3-D spatial agent-based simulation of tumor spheroid response to a chemotherapy to test multi-substrate representation[1]	Vectors of substrates
Chaste_ABM	3-D tumor spheroid simulation[3,4]	Additional cell cycle models
DCIS_ABM	previously published agent-based model outputs, translated to prototype MultiCellXML to current standard[6]	
GBM	outputs from a continuum (cell density) simulation of glioblastoma multiforme[7,8].	Clinical annotations
mouse_liver	3-D segmented liver pathology[9,10]	Refinements to vascular elements, representation of pixel-level data
BC_pathology	nuclear segmentation and quantitation applied to previously published breast cancer pathology images[17]	Clinical annotations

Table 3. Summary of digital snapshots in the library.

data element	a specific measurement (e.g., the radius of a cell nucleus); the smallest unit of data in MultiCellDS
XML	An extensible markup language (similar to HTML used for webpages), which can order data element hierarchically to match their relationships in biology
XML schema	A template for XML data, which describes the allowed data elements and how they can be arranged
metadata	Data about the data: extra information such as units, scale, uncertainty, or provenance
ontology	A controlled dictionary of allowed terms
OWL	A standardized ontology file, in the "web ontology language" format
provenance	The history of who created data, who revised it, who maintains it, and where it is published.
MultiCellDS	multicellular data standard
MultiCellDB	MultiCellDS database
MultiCellXML	MultiCellDS data, written in an XML format
phenotype dataset	A collection of cell phenotype and other measurements, in a single microenvironmental context
digital cell line	A collection of phenotype datasets and key metadata for a single biological cell line or type
digital snapshot	A readout of all cells, their phenotypes, and the microenvironment at a single time
collection	A logical grouping of one or more digital cell lines, digital snapshots, collections, or a combination of these

Box 1. Summary of digital snapshots in the library.