CellPress

An open data ecosystem for cell migration research

Paola Masuzzo[1,2], Lennart Martens[1,2], and The 2014 Cell Migration Workshop Participants[3]

[1] Department of Medical Protein Research, VIB, A. Baertsoenkaai 3, 9000 Ghent, Belgium
[2] Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, 9000 Ghent, Belgium
[3] See Contributing Authors section at the end of this paper

Cell migration research has recently become both a high content and a high throughput field thanks to technological, computational, and methodological advances. Simultaneously, however, urgent bioinformatics needs regarding data management, standardization, and dissemination have emerged. To address these concerns, we propose to establish an open data ecosystem for cell migration research.

Where is the cell migration field migrating to?

Cell migration is crucial in biological processes such as morphogenesis, immune surveillance, wound healing, and cancer metastasis [1]. Diverse biological models have been developed to reflect the range of molecular and physiological events involved in cell migration (see Figure S1 in the supplementary material online). Furthermore, technology has been an important driver for innovation in cell migration research. For example, the evolution of light microscopy from bright field to confocal, two photon, light sheet, and superresolution fluorescence microscopy has enabled the development of complex experimental systems, progressing from 2D cell migration assays to 2.5D and 3D (see Glossary) approaches [2] (see Table S1 in the supplementary material online).

While analyses on 2D substrates have led to essential insight into the cellular motility machinery, 3D environments are essential for understanding their physiological context, and have recently provided novel knowledge regarding invasive behaviour [3]. Although these *in vitro* assays are clearly valuable, deeper insight into cell migration can only be obtained through *in vivo* approaches. Such assays have been enabled through live cell microscopy to visualize moving cells in their native surroundings, revealing previously unsuspected feedback mechanisms [4]. Moreover, results in high content and high throughput microscopy have established the importance of quantitative analysis for systems biology and drug discovery [5].

A key remaining challenge is to understand how the function and signalling of organelles is coordinated and integrated within cells and tissues. Cell migration is the product of complex processes operating at different scales, and could be investigated using a systems microscopy approach [6], combining image analysis at different resolutions with data mining, multivariate statistics, and modelling.

These advances in techniques and biological models have been supported by dedicated efforts in bioinformatics and computational biology (see Table S1 in the supplementary material online). Algorithms and tools have been developed for tracking cells using time lapse images [7], and for processing and visualizing large sets of complex image data (http://jcb-dataviewer.rupress.org). The computational approaches in the field extend to *in silico* modelling of cell migration and invasion, especially in tumour development and progression [8]. Advances in the field have thus been built on a combination of novel analytical approaches, dedicated software tools and algorithms, and predictive theoretical models.

Taking on the challenges: an open data ecosystem for cell migration

Even though the cell migration field has embraced computational models as a means to integrate and interpret experiments, a key missing element is the global iterative connection between experimental data and computational approaches. This connection requires an open and free data ecosystem, where standardized and documented results of cell migration research can be shared and consulted within a central location, as exemplified in Figure 1. Building such an ecosystem will require several interdigitated and essential developments. A public, centralized repository constitutes the major component; however, it is only viable if supported by standard formats for the stored data and metadata. Furthermore, each data set in the repository should conform to minimum reporting requirements that ensure consistent annotation (see Table S1 in the supplementary material online). The following sections describe each of these aspects in more detail.

Data and metadata standardization

Minimum reporting requirements

To be reusable, an experimental data set needs accompanying metadata, describing both biological and

Glossary

2D: two-dimensional.
2.5D: two-and-a-half-dimensional.
3D: three-dimensional.
CMC: cell migration consortium.
CMG: cell migration gateway.
CV: controlled vocabulary.
OME: open microscopy environment.

Corresponding author: Martens, L. (lennart.martens@vib-ugent.be).
Keywords: cell migration; bioinformatics; standardization; meta-analysis data ecosystem.

0962-8924/
 http://dx.doi.org/10.1016/j.tcb.2014.11.005

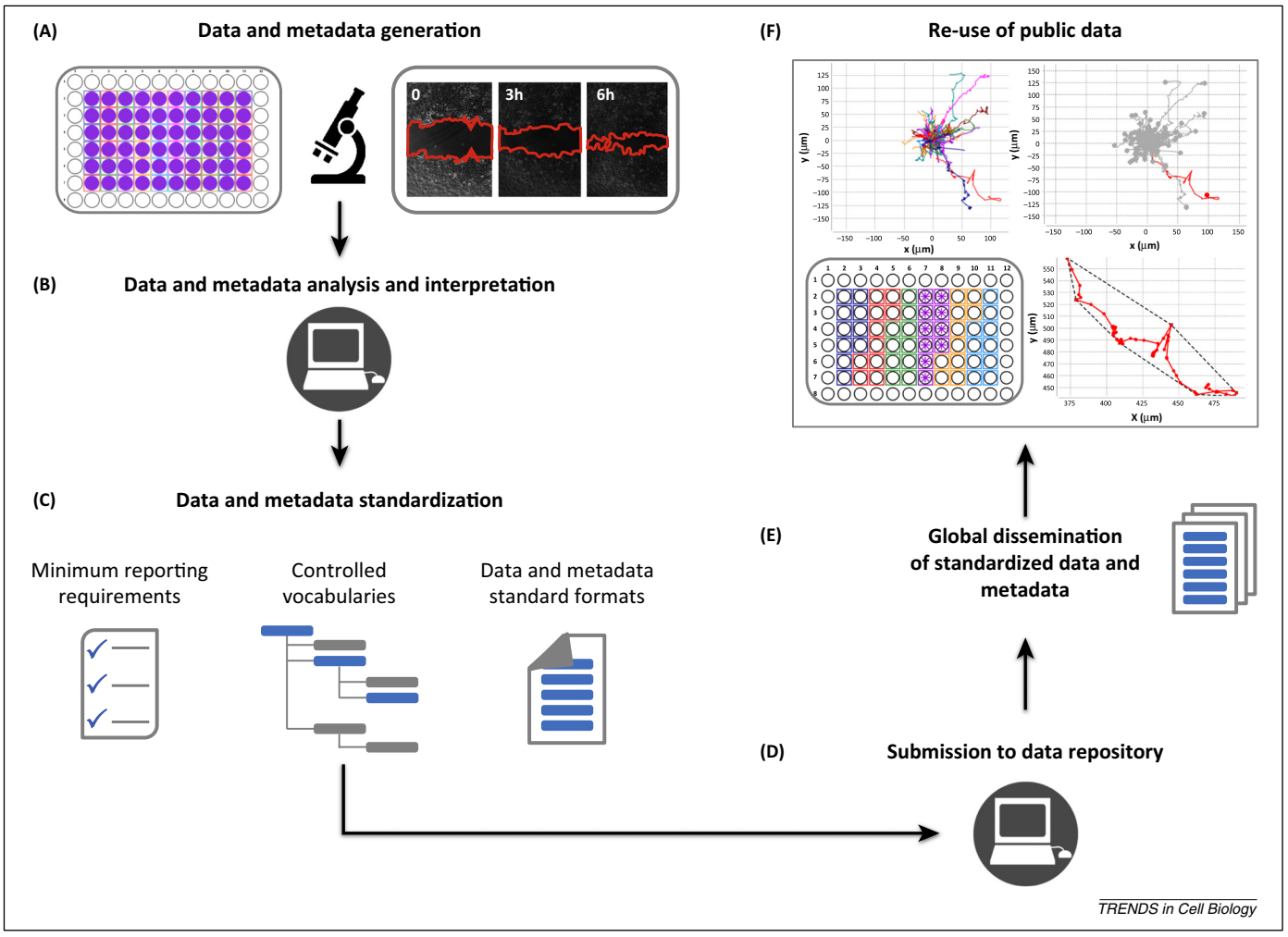

Figure 1. An example of an experimental workflow in the open data ecosystem. **(A)** Data and metadata associated with an experiment are generated. **(B)** Software is used to analyse and interpret the resulting data and associated metadata. **(C)** The collected data are formatted and reported in the relevant standards to enable data and metadata reproduction, verification, and exchange: minimum reporting requirements specify the core information to be supplied through the software tool; controlled vocabularies (CVs) are used to unambiguously annotate such units of information; and the data are exported using data and metadata standard formats. A fully standards compliant cell migration data set is ready for **(D)** submission to, and **(E)** subsequent dissemination from, a global data repository. **(F)** The open data sharing ecosystem will enable the re-use of public cell migration data, including multiscale and meta-scale analyses across large scale experiments, ultimately unlocking new knowledge in the field.

methodological context. Community-wide minimum reporting requirements have, therefore, been created in many fields, for example, for proteomics [9] and, of direct interest to cell migration, for cell perturbation experiments (http://miaca.sourceforge.net/). The global harmonization of such field-specific minimum information checklists is pursued by the BioSharing project (http://biosharing.org/).

The existing requirements can serve as a starting point to build a specific checklist for *in vitro* cell migration experiments. A tentative example of what such a list could look like is shown in Table 1: example information is provided about experimental modules and submodules, from sample preparation over image acquisition and analysis, to downstream data analysis, and laboratory metadata. A second iteration can then extend this to *in vivo* studies, which will be more challenging.

Controlled vocabularies

Minimum reporting requirements specify which information should be reported, but not yet how this information should be conveyed. The use of a common terminology thus becomes important, typically taking the form of a controlled vocabulary (CV). Again, proteomics provides an example of such a CV for the unique and unambiguous, yet detailed semantic annotation of (meta-)data [10]. Existing CVs that can be reused for cell migration experiments include the Cell Ontology [11] and the Cellular Microscopy Phenotype Ontology (http://www.ebi.ac.uk/cmpo/).

Standard data and metadata formats

When minimum reporting requirements are coupled to CVs, data and metadata can be conveyed in an unambiguous and well documented form. However, one more element is needed for successful standardization: the adoption of standard data formats. As in any data rich field, software tools are continuously applied in cell migration research to process and analyse data. However, such software can only read data presented in known formats, usually dictated by instrument vendors, and therefore implying that data can only be read by other researchers if they have access to the same instrument. Moreover, such proprietary data formats also suffer from data rot [12]. These issues can be resolved through community standard, open data formats, where considerable work