Del Sol, A., Thiesen, H. J., Imitola, J., & Carazo Salas, R. E. (2017). Big-Data-Driven Stem Cell Science and Tissue Engineering: Vision and Unique Opportunities. *Cell Stem Cell*, *20*(2), 157-160. https://doi.org/10.1016/j.stem.2017.01.006

Big Data-Driven Stem Cell Science and Tissue Engineering: Vision and Unique Opportunities

Antonio Del Sol [1*], Hans J. Thiesen [2], Jaime Imitola [3], Rafael E. Carazo Salas [4,5*]

[1] Computational Biology Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6, Avenue du Swing, Belvaux 4367, Luxembourg
[2] Institute of Immunology, Universitätsmedizin Rostock, University of Rostock, Rostock, Germany
[3] Laboratory for Neural Stem Cells and Functional Neurogenetics, Departments of Neurology and Neuroscience, and Molecular Biology and Cancer Genetic Program, The James Comprehensive Cancer Hospital, Columbus, Ohio, The Ohio State University Wexner Medical Center, USA.
[4] Pharmacology Department and Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1PD, UK
[5] Current address: School of Cell and Molecular Medicine, University of Bristol, Biomedical Sciences Building, University Walk, Bristol BS8 1TD, UK

* Correspondence:
antonio.delsol@uni.lu
and
cre20@cam.ac.uk

Achieving the promises of stem cell science to generate precise disease models and 'designer' cell samples for personalized therapeutics will require harnessing pheno-genotypic cell-level data quantitatively and predictively in the lab and clinic. Those requirements could be met by developing a Big Data-driven stem cell science strategy and community.

Why Big Data-driven stem cell science?

Stem cell science has seen a revolution in the past decade. With the advent of human pluripotent stem cell (hPSC) technologies 10 years ago and more recently of CRISPR/Cas9 genetic engineering and organoid technologies, the scope of the stem cell field has expanded significantly to encompass Biomimetic Engineering, which promises to provide in the not-so-far future 'designer' cells, and tissues and organs for the precise study and personalized treatment of countless diseases, ranging from heart and liver failure, to sickle cell anemia, macular degeneration and Alzheimer's disease.

To achieve that unique goal, stem cell science will need to harness pheno-genotypic 'omics' Big Data, that is genomic, transcriptomic, proteomic, epigenomic, microscopic, metabolomic and other such information, to learn how to efficiently, specifically and safely produce cells that best match those of each intended target patient.

We propose that this can be best attained by defining a collaborative strategy for data-driven stem cell science, specifically to: a) coordinate efforts of the community to generate and share high-quality stem cell 'omics' reference datasets, which can be used for defining qualitatively and quantitatively different cell/tissue/organ types and how those differ between individuals; and b) engage with the computational biology community, which is currently under-represented in the stem cell field and will be pivotal in developing the technologies needed both to leverage and integrate biological 'omics' Big Data and to establish how much pheno-genomic information is enough to consider a designer cell or tissue as acceptable for responsible therapeutic use.

The peculiarities of Big Data-driven stem cell science

Contrary to communities like the cancer field, which exploits pheno-genotypic data to precisely identify cancer cells and tissues and use that knowledge to find new ways to destroy them, a major focus of the stem cell field is instead to leverage pheno-genotypic information to engineer cells/tissues that are as close as possible to a desired cell/tissue type for a specific target individual. This has become theoretically possible thanks to a number of technological breakthroughs, mostly: hPSC technologies, including human embryonic stem cell (hESC) and particularly induced pluripotent stem cell (hiPSC) technologies and the development of direct reprogramming, directed differentiation and trans-differentiation techniques; CRISPR/Cas9 genetic engineering; and organoid culturing methods using hESC, hiPSC and adult stem cells. Together those breakthroughs have made it conceivable that we will be able not too far in the future to generate *in vitro* designer cells, tissues and organs of any type, at will and at scale (Hockemeyer and Jaenisch, 2016).

This peculiar - designer - focus of the field in turn makes the requirements of data-driven stem cell science very different from those of other fields at multiple levels (**Figure 1**). For cells, the goal is to produce 'engineered' cells that match the identity (Cahan et al., 2014) of 'intended' cells at the transcriptional, epigenetic and other phenotypic levels. For tissues and organ(oid)s, the aim is to produce tissues/organs with similar competence, proliferation, stratification (spatio-temporal organization in 2D/3D) and functional properties (e.g. contraction for cardiomyocytes, action potential for neuronal rosettes) to those of the intended tissues/organs to be replaced. For samples intended for real-life clinical use in a patient, one would aim to match as best as possible the patient's pheno-genotypic profile in the case of heterologous cells/tissues/organs and potentially match the intended niches that will host the samples within the patient.

We argue that quantitative, omics-driven approaches can satisfy these needs at each level by: a) precisely defining different cellular identities; b) using that information to learn how to make engineered cells/tissues that best match their intended counterparts; and c) establishing quality standards for acceptable use of stem cell-derived products for personalized therapeutics and building on standards currently being put in place for their clinical evaluation (for example (Kawamata et al., 2015)).

We think it is crucial at this point in time that those approaches be openly discussed and agreed together by the stem cell community, for instance at dedicated meeting sessions or task forces, to potentiate future progress in the field. As we discuss below, we propose that the community should decide together on the quantitative data-driven standards to be used at each level, how to revise them as technology and data science advance, and the appropriate coverage (cell/tissue/organ types), resolution (types of 'omics data) and depth (quantity) of pheno-genomic data that are needed for academic versus clinical versus industrial applications.

Generating and sharing integrable stem cell pheno-genomic data at the community level

An increasing number of initiatives have recognized that there is a pressing need to generate high-quality, curated pheno-genomic stem cell datasets for use by the community.

Examples of such initiatives providing different levels of pheno-genomic information for stem cells are the ENCODE Project Consortium (https://www.encodeproject.org/), the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/), the EBI Gene Expression Atlas (https://www.ebi.ac.uk/gxa/home), NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and the Human Pluripotent Stem Cell Registry (http://hpscreg.eu/). These initiatives were historically isolated and generated their datasets independently of one another. This has inevitably meant that there was absence of standardization in the datasets produced, that there might have been partial duplication of some efforts that could otherwise have been invested in increasing the coverage of further cell/tissue/organ types, and that the potential added value coming from combining some of the information contained in those datasets could not be properly exploited. However, as the initiatives focused on mostly different stem cell types/sources, and

because they were not intended to be integrated or comply with the current requirements of the stem cell field, their isolation did not represent a hindrance.

The recently released iPS cell pheno-genomic datasets of the HipSci Consortium (http://www.hipsci.org/, (Kilpinen et al., 2016)) and the Progenitor Cell Biology Consortium (https://www.synapse.org/#!Synapse:syn1773109/wiki/54962, (Salomonis et al., 2016)) instead exemplify what we mean. Both datasets represent truly invaluable resources for the community and both are the result of years of top quality work. Nevertheless, in their current form their data (for example, mRNA or methylation array) cannot be directly compared or indeed integrated, which severely limits potential joint added value that could have arisen from combining them.

We anticipate that from now on there will need to be an increasing effort to coordinate the generation of thorough pheno-genomic datasets for stem cells, intended cell types and engineered cell types, to facilitate the standardization, comparison and integration of those datasets.

This will be important for many reasons. First, this integrated approach will enable researchers to compare and combine detailed data coming from similar cell types (for instance, coming from hiPS-derived cardiomyocytes generated in different studies or using different cell sources within a study), in order to obtain a more 'precise' description of what constitutes a given type of cell, tissue, organ, niche or individual. Secondly, integrating such data will help establish what constitutes the 'normal' healthy baseline state of a given cell/tissue type. Indeed, understanding the baseline molecular phenotypes and the dynamics of progenitor and differentiated cells in each lineage/tissue is likely to be essential for the rational development of cell therapy. Initiatives such as the Human Cell Atlas (https://www.broadinstitute.org/research-highlights-human-cell-atlas) with a focus on single cell analysis might be key to help gain those critical insights. Thirdly and finally, coordinating the generation of stem cell pheno-genomic datasets will be crucial for clarifying the impact pheno–genotypic differences have on the production of different cell/tissue types and, in this manner, for gaining a better understanding of how to engineer cells and tissues 'personalized' to best match a given individual for therapeutic use.

We propose therefore to establish a community strategy to generate, where possible standardize, and openly share stem cell pheno-genomic datasets in a coordinated manner, with the goal to facilitate their further comparison and integration. A similar type of strategy has been recently proposed for the cancer community by the Cancer Moonshot Task Force report (https://www.whitehouse.gov/sites/default/files/docs/final-whcmtf-report-1012161.pdf).

In particular we suggest to:
(1) Identify at the community level which types of stem cell biological Big Data are suitable to become reference for greater community use and hence which standard datasets should be available or generated;

(2) Agree on the types of cells/tissues to be included, and whether the data should come from single cells or bulk cell population measurements as well as single timepoint and/or longitudinal time-lapse datasets. Indeed, stem cell differentiation is rarely homogenous and stem cell phenotypic subsets usually reside within populations whose average state may not reveal their properties. Hence identifying ways to share, compare and integrate pheno-genomic data (preferably from the same experimental system) from large numbers of single cells that capture cellular heterogeneity might be key. This includes multi-parametric, time-resolved phenomic data on dynamical processes (like cell cycle, shape, polarity, migration or death, or changes in transcription factor status), which could be derived from single-cells or cell clones *in vitro* by time-lapse microscopy analysis and might be key to better define different cell/tissue states and how cells/tissues transition between states;
(3) Establish the (ethnic, geographic, etc) diversity of patients/individuals to be included;
(4) Define how to handle datasets that are medically related, particularly if and how they can be made available to the community in a way that is safe and respects privacy of the individuals; and
(5) Agree on standards for size, depth, quality and annotation of datasets, and for sharing those datasets.

We think this could be best achieved by actively involving bioinformaticians and computational biologists in those discussions, to establish all together the quantity and quality of each type of data needed to generate integrated, meaningful analyses.

Technologies needed to leverage and integrate stem cell biological Big Data
Of course, data is only useful if it improves the capacity to achieve intended goals.

For stem cells, the goals are to precisely define different cellular identities and exploit that information both to engineer matching cells/tissues for basic and clinical research and to provide quality standards for prospective personalized therapeutics 'products'.

In the case of cellular identities, the stem cell community has been increasingly making use of biological Big Data analytics and integration strategies to better define cell types or states and the differences among cell types/states (for a recent example from one of our groups see (Okawa et al., 2016)). Popular approaches have included Principal Component Analysis, Correlation Analysis, and Clustering (hierarchical, K-means, etc), however recently more sophisticated mathematical and computational methods, including Machine Learning methods, have begun being used to address questions about cellular state and identity in a more specific, quantitative, data-driven and predictive manner. For instance, the online frameworks CellNet (http://cellnet.hms.harvard.edu/) and Mogrify (http://www.mogrify.net/) use microarray data and network mathematics to quantitatively define and compare cell types/states and predict candidate transcription factors to change between cellular states. We foresee that predictive methods will become pivotal in tackling the emerging application requirements of the stem cell field.

In particular, Artificial Intelligence approaches able to integrate high-dimensional data and make it highly predictive with little prior knowledge or supervision, like Deep

Learning ((LeCun et al., 2015, Mamoshina et al., 2016)), and approaches able to cumulatively port predictive knowledge gained from existing datasets to new datasets, like Reinforcement Learning or Transfer Learning ((Kandaswamy et al., 2016, LeCun et al., 2015)), might be best suited for the therapeutic ambitions and requirements of data-driven stem cell science.

Those techniques will be key to establish quantitative 'Turing Test'-like approaches (Figure 1; (Cronin et al., 2006)) allowing to establish how much information is sufficient and what statistical standards and accuracy are needed to ascertain when an engineered cell/tissue is indistinguishable from a desired target cell/tissue, so as to fulfill the corresponding basic and clinical research requirements.

Crucially, we think those approaches will be pivotal to better define when engineered cells/tissues can be safely used for therapeutic and biotechnological use, particularly to establish quantitative production standards demonstrating the cells/tissues' safety (non-tumourigenicity) and predictive therapeutic potential, in Regenerative Medicine applications.

We propose to start as a community a discussion about which approaches will be needed to fulfill the future quantitative standards that we require, and to actively engage with and channel into that discussion the Machine Learning and Artificial Intelligence communities to help us develop those predictive approaches.

Consolidating the computational stem cell biology community

Finally, the key to fulfill the vision and opportunities of Big Data-driven stem cell science will be to establish and consolidate a dedicated computational stem cell biology community.

Up till now there has been overall a general disconnect between the stem cell community and that of computational biology. As an illustration of this disconnect, 20 out of 22 stem cell conferences advertised on the ISSCR events page for 2016/2017 (http://www.isscr.org/home/events/stem-cell-meetings) all lack a bio-computational component, with only two recent exceptions having such a discussion held in a break-out session (for example, http://www.isscr.org/home/internationalsymposia/previous-events/dresden-2016/home). Reciprocally, in conferences on data-driven biomedicine there has traditionally been no agenda vis à vis stem cells, contributing to the divide.

We suggest from now to include dedicated stem cell computational biology sessions in the major stem cell community meetings (for example, the Annual ISSCR Meetings), to pro-actively channel and integrate into our community the bio-computational field given that this field will play a fundamental role in precision and therapeutic stem cell science. A session could contain talks and discussions spanning the following thematic areas: large scale data acquisition; data repositories; online repositories; development of standard tools for analytics; development of novel quantitative methods for precision research and therapeutics; data sharing strategies; data integration strategies; and interaction with the clinical community and clinical data.

In conclusion, we propose that a better definition of an agenda for quantitative Big Data-driven stem cell science by the community (particularly thought leaders and decision makers) and the concomitant development of a strong computational stem cell biology community oriented toward our community's unique research and clinical/industrial requirements will be pivotal to help make the stem cell field quantitative, predictive and therapeutically-suitable and make the promise of stem cell therapeutics a reality.

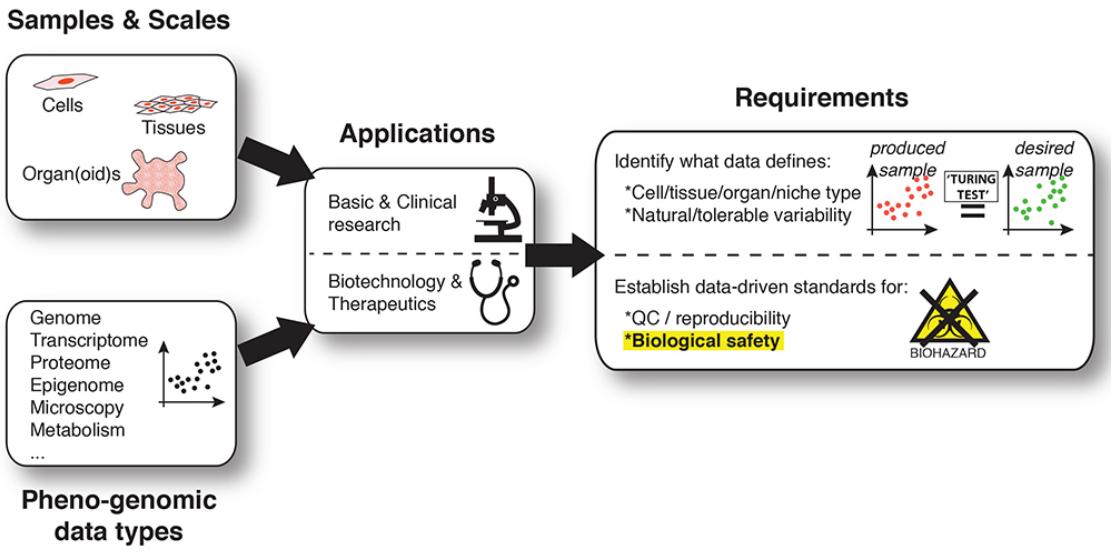

Figure 1. Unique data-driven requirements of stem cell science. The diagram depicts the diverse and unique scope of current stem cell science and corresponding data-driven requirements: scales of cell samples, data types accessible, applications and main data-driven requirements and questions that should be addressed to establish quantitatively and responsibly the adequacy of designer cells/tissues/organs for Regenerative Medicine applications.

CAHAN, P., MORRIS, S. A., COLLINS, J. J. & DALEY, G. Q. 2014. Defining cellular identity through network biology. *Cell Cycle,* 13**,** 3313-3314.

CRONIN, L., KRASNOGOR, N., DAVIS, B. G., ALEXANDER, C., ROBERTSON, N., STEINKE, J. H. G., SCHROEDER, S. L. M., KHLOBYSTOV, A. N., COOPER, G., GARDNER, P. M., SIEPMANN, P., WHITAKER, B. J. & MARSH, D. 2006. The imitation game - a computational chemical approach to recognizing life. *Nature Biotechnology,* 24**,** 1203-1206.

HOCKEMEYER, D. & JAENISCH, R. 2016. Induced Pluripotent Stem Cells Meet Genome Editing. *Cell Stem Cell,* 18**,** 573-586.

KANDASWAMY, C., SILVA, L. M., ALEXANDRE, L. A. & SANTOS, J. M. 2016. High-Content Analysis of Breast Cancer Using Single-Cell Deep Transfer Learning. *Journal of Biomolecular Screening,* 21**,** 252-259.

KAWAMATA, S., KANEMURA, H., SAKAI, N., TAKAHASHI, M. & GO, M. J. 2015. Design of a Tumorigenicity Test for Induced Pluripotent Stem Cell (iPSC)-Derived Cell Products. *Journal of Clinical Medicine,* 4**,** 159-171.

KILPINEN, H., GONCALVES, A., LEHA, A., AFZAL, V., ASHFORD, S., BALA, S., BENSADDEK, D., CASALE, F. P., CULLEY, O., DANACEK, P., FAULCONBRIDGE, A., HARRISON, P., MCCARTHY, D., MCCARTHY, S. A., MELECKYTE, R., MEMARI, Y., MOENS, N., SOARES, F., STREETER, I., AGU, C. A., ALDERTON, A., NELSON, R., HARPER, S., PATEL, M., CLARKE, L., HALAI, R., KIRTON, C. M., KOLB-KOKOCINSKI, A., BEALES, P., BIRNEY, E., DANOVI, D., LAMOND, A. I., OUWEHAND, W. H., VALLIER, L., WATT, F. M., DURBIN, R., STEGLE, O. & GAFFNEY, D. J. 2016. Common genetic variation drives molecular heterogeneity in human iPSCs. *bioRxiv*.

LECUN, Y., BENGIO, Y. & HINTON, G. 2015. Deep learning. *Nature,* 521**,** 436-444.

MAMOSHINA, P., VIEIRA, A., PUTIN, E. & ZHAVORONKOV, A. 2016. Applications of Deep Learning in Biomedicine. *Molecular Pharmaceutics,* 13**,** 1445-1454.

OKAWA, S., NICKLAS, S., ZICKENROTT, S., SCHWAMBORN, JENS C. & DEL SOL, A. 2016. A Generalized Gene-Regulatory Network Model of Stem Cell Differentiation for Predicting Lineage Specifiers. *Stem Cell Reports,* 7**,** 307-315.

SALOMONIS, N., DEXHEIMER, P. J., OMBERG, L., SCHROLL, R., BUSH, S., HUO, J., SCHRIML, L., SUI, S. H., KEDDACHE, M., MAYHEW, C., SHANMUKHAPPA, S. K., WELLS, J., DAILY, K., HUBLER, S., WANG, Y. L., ZAMBIDIS, E., MARGOLIN, A., HIDE, W., HATZOPOULOS, A. K., MALIK, P., CANCELAS, J. A., ARONOW, B. J. & LUTZKO, C. 2016. Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium. *Stem Cell Reports,* 7**,** 110-125.