Published in final edited form as:
Cell Stem Cell. 2014 April 3; 14(4): 427–430. doi:10.1016/j.stem.2014.03.014.

Identifiability and Privacy in Pluripotent Stem Cell Research

Rosario Isasi[1,*], **Peter W. Andrews**[2], **Jay M. Baltz**[3], **Annelien L. Bredenoord**[4], **Paul Burton**[5], **Ing-Ming Chiu**[6], **Sara Chandros Hull**[7], **Ji-Won Jung**[8], **Andreas Kurtz**[9,10], **Geoffrey Lomax**[11], **Tenneille Ludwig**[12], **Michael McDonald**[13], **Clive Morris**[14], **Huck Hui Ng**[15], **Heather Rooke**[16], **Alka Sharma**[17], **Glyn N. Stacey**[18], **Clare Williams**[19], **Fanyi Zeng**[20], and **Bartha Maria Knoppers**[1]

[1]Centre of Genomics and Policy, Faculty of Medicine, Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada [2]Department of Biomedical Science, University of Sheffield, Sheffield S10 2TN, UK [3]Ottawa Hospital Research Institute and Department of Obstetrics and Gynecology, Faculty of Medicine, University of Ottawa, Ottawa, ON K1H 8L6, Canada [4]Julius Center, Department of Medical Ethics, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands [5]School of Social and Community Medicine, University of Bristol, Bristol BS8 1TH, UK [6]Regenerative Medicine Research, ICSM, National Health Research Institutes, Jhunan, Miaoli 35053, Taiwan [7]NHGRI Bioethics Core, National Institutes of Health, Bethesda, MD 20892-1156, USA [8]Division of Intractable Diseases, Center for Biomedical Sciences, National Institute of Health and Korea Centers for Disease Control and Prevention, Chungcheongbuk-do 363-951, Republic of Korea [9]Berlin-Brandenburg Center for Regenerative Therapies, Berlin 13353, Germany [10]Seoul National University, College of Veterinary Medicine and Research Institute for Veterinary Science, Seoul 151-742, Republic of Korea [11]California Institute for Regenerative Medicine, San Francisco, CA 94107, USA [12]WiCell Research Institute, Madison, WI 53726, USA [13]W. Maurice Young Centre for Applied Ethics, School of Population and Public Health, University of British Columbia, Vancouver, BC V6T 1Z2, Canada [14]National Health & Medical Research Council, Canberra ACT 2601, Australia [15]A*STAR, Genome Institute of Singapore, Singapore 138672, Singapore [16]International Society for Stem Cell Research, Skokie, IL 60077, USA [17]Medical Biotechnology Division, Department of Biotechnology, Ministry of Science & Technology, Government of India, New Delhi 110-003, India [18]National Institute for Biological Standards and Control: A Centre of the MHRA, South Mimms, Hertfordshire UB8 3PH, UK [19]Department of Sociology & Communications, School of Social Sciences, Brunel University, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK [20]Shanghai Institute of Medical Genetics, Shanghai Stem Cell Institute, Shanghai 200025, People's Republic of China

Abstract

Data sharing is an essential element of research; however, recent scientific and social developments have challenged conventional methods for protecting privacy. Here we provide guidance for determining data sharing thresholds for human pluripotent stem cell research aimed at a wide range of stakeholders, including research consortia, biorepositories, policy-makers, and funders.

[*]Correspondence: rosario.isasi@mcgill.ca.

The discovery of technologies to generate induced pluripotent stem cell (iPSC) lines and the corresponding derivation of large numbers of these lines for research and potential therapeutic use have resulted in a rejuvenated interest in biorepositories (McKernan and Watt, 2013; Stacey et al., 2013). Biorepositories are vital infrastructures providing primary material (primary samples, cell lines, and associated data) for research and clinical translation. Today, biorepositories serve also as the primary resource for authenticated, quality controlled, and ethically sourced human pluripotent stem cell (hPSC) lines. Robust banking networks now enable global access to well-characterized and traceable hPSCs, an essential prerequisite for scientific reproducibility (Stacey et al., 2013). The availability of such resources presents a wide range of therapeutic opportunities; however, sharing them also comes with an attendant responsibility to protect donors' or research participants' (hereinafter “participants”) privacy.

These competing factors require striking a delicate balance between the amount and quality of data collected and the precautions taken when sharing such information. Comprehensive data curation is important because cell-line misidentification continues to be a pervasive problem, undermining the scope and authenticity of research findings. In addition, well-annotated genomic and epigenomic data, and participants' phenotypic and demographic data, facilitates disease modeling and drug development and contributes to the understanding of genetic variation and its role in normal cell behavior. Next generation sequencing (NGS) technologies combined with bioinformatic data systems enable data analysisona wide range of participants, facilitating the translation of cell-based-therapies (Kreiner and Irion, 2013; McKernan and Watt, 2013).

In this Forum,we discuss the challenges of establishing thresholds for sharing and publishing individual/summary data associated with hPSC research. We review the ensuing scientific, socioethical, and legal implications and propose a framework with criteria for data sharing policies. Our recommendations are directed at a wide range of stakeholders.

Data Sharing, Privacy, and Reidentifiability

Fundamental scientific data can be perceived as a community resource. Data sharing constitutes an ethical and scientific imperative that is recognized by international funders and scientific organizations across disciplines and jurisdictions (Knoppers, 2010; Kaye, 2012). This imperative is underpinned by the principles of reciprocity, solidarity, and respect for all stakeholders. Data sharing is envisaged as a tripartite responsibility of data producers, users, and funders (Isasi et al., 2012; Knoppers et al., 2011). Scientific integrity and progress are dependent not only on the sharing of raw data between researchers, but also on the ability to widely disseminate research findings. In turn, public trust is earned and maintained by responsible stewardship. The latter entails protecting—and possibly also promoting—the interests of participants while advancing societal benefits. Moreover, trust requires respecting divergent interests by balancing benefits and risks in a proportionate and appropriate manner (Rodriguez et al., 2013).

Several scientific and social developments are prompting reconsideration of how the imperative of data sharing is conceptualized and implemented. The decreasing costs and increasing accessibility of NGS and cloud computing, along with the growing volume, richness, and complexity of genomic information available, are challenging individual privacy and the traditional methods designed to manage and secure such data (e.g., coding and anonymization). These factors, together with reports of the ease of reidentification in the scientific literature and popular press, contribute to changing public attitudes on the meaning of individual privacy and attendant expectations about the fiduciary duties of data stewards (Kaye, 2012; Rodriguez et al., 2013).

Empirical studies to assess participants' data sharing decisions and attitudes demonstrate that regarding clinical and genetics research, participants are overall “health informational altruists.” Such studies are reassuring because they suggest that an inability to guarantee privacy may not deter individuals from participating in research (Rodriguez et al., 2013). However, there is also a need to consider mitigating actions to ensure that participants trust in science. For example, participants often wish to be involved in decision-making and have concerns about governance mechanisms safeguarding privacy. In addition, these studies show that participants' privacy-utility trade-off decisions vary in real versus hypothetical scenarios (Kaye, 2012). With hPSC research specifically, there is emerging evidence that participants broadly support data sharing even while maintaining privacy concerns. Further research is needed to systematically assess participants' views (Dasgupta et al., 2014).

We are at a crucial juncture where novel statistical methods and associated tools allow the drawing of inferences, possibly revealing the identity of individual participants in biomedical research. Genomic information is both intrinsically self-identifying and a source of familial information. A recent study demonstrated that re-identification is possible even in the absence of a reference sample (Gymrek et al., 2013). Several genomic studies also demonstrated a wide range of scenarios in which reidentifying participants in biomedical research could be possible by triangulating multiple publicly available data sources (e.g., census and genealogy data, obituaries, voter registries, etc.). It has been established that relying on as few as 75 individual (statistically independent) SNP loci could enable unique individual identifiability (Gymrek et al., 2013; Rodriguez et al., 2013).

Reidentification is the ability of protected data to be traced back to a participant. It can occur directly or indirectly, deliberately or unintentionally, and by different means: (1) directly, by matching genomic data against a reference genotype; (2) by deduction, or by linking to nongenetic databases (e.g., health care, forensic, administrative, genealogical, etc.) and matching it to genotype and other associated data (e.g., gender, age, disease status, etc.); and (3), by inference, by profiling genomic data from DNA analysis (e.g., gender, blood type, etc.). Consequently, individual identifiability is currently present at incremental levels from overtly identifiable to potentially identifiable (Rodriguez et al., 2013; Gymrek et al., 2013; Kaye, 2012).

While the generalizability of the above-mentioned methods and tools continues to be evaluated, and evidence-based risk reassessments continue to be debated, it is clear that the concepts of identifiability and privacy are shifting, as are the expectations of stakeholders. For a proportional and realistic risk assessment, due consideration should be given not only to the existence of multiple data resources, potential data users, and malicious intruders, but also to different data environments as a whole, which extend well beyond the research context. Privacy risk assessments should also be situated in a society in which social media and direct-to-consumer genetic testing are omnipresent (Knoppers, 2010). In this manner, individuals are broadly and openly sharing their personal information, genomic or otherwise, as well as their family members' information, either directly or by association. These factors increase the likelihood of participant reidentification by expanding the range of data resources publicly available that can be used in combination with other data sources to reidentify individuals. They further create vulnerabilities in governance mechanisms, decrease the effectiveness of data security measures, challenge the protections for privacy and confidentiality, and thereby provide an opportunity for participant reidentification (Gymrek et al., 2013; Rodriguez et al., 2013). For these reasons, relying solely on traditional methods based on informed consent and data coding or anonymization (irreversibly stripping of identifiers) is naive and insufficient to protect participants' privacy (Kaye, 2012). More sophisticated security measures, in combination with sanctions for deliberate breaches of confidentiality, are required to keep pace with technological developments (Knoppers et al., 2011).

A pivotal concern regarding identifiability is the potential for personal and health information to be associated with a specific individual and the possible harms of discrimination (e.g., in employment or insurance), stigmatization, stress, and anxiety (Kaye, 2012). These harms need not be confined to the individual but could also be extended to a community or subpopulation to which the participant belongs (based on disease condition, ethnicity, or familial relations). Needless to say, unintended or deliberate misuse and disclosure of personal information due to participant reidentifiability breaches the trust established between researchers and participants. Therefore, risks and harms are not restricted solely to participants but are also present for data stewards, researchers, and the entire scientific enterprise (Rodriguez et al., 2013; Knoppers et al., 2011).

It should also be emphasized, however, that at the present time, concerns about the reidentification of genomic data in the research context are largely hypothetical. There are no known/published reports of breaches of confidentiality resulting in actual harm to participants in genetic research. Published examples using statistical methods to reidentify genomic data have been proofs of concept rather than malicious uses of data (Gymrek et al., 2013; Rodriguez et al., 2013).

Scientific Considerations for hPSC Line Derivation

Given that an hESC line reflects the contributions of two genetically different individuals, genetic/genotype data arising from an hESC line itself is unique to the embryo/cell line and not directly attributable to any individual donor. For this reason, the possibility of donor reidentification based *solely* upon the genotype of an hESC line remains extremely remote. However, while hESC-associated data would not correspond directly to the genotype of the individual donor or donors, the information that can be gleaned using diverse molecular analyses could have medical and social significance for the donors and related individuals. Moreover, in some cases the interpretations of certain genetic data derived from numerous loci (e.g., ethnicity), combined with the laboratory of origin or partial genotype information for a putative donor or donors, could be sufficient for the donors to identify themselves or be identified by others by triangulation with public information (Knoppers et al., 2011; Isasi et al., 2012).

In contrast to hESC lines, iPSCs contain donor-specific DNA. While the gene insertion and reprogramming process results in minor changes to the DNA (such as changes in methylation patterns) the genetic/genomic data arising in this context remains virtually identical to that of the donor. Consideration should be given to circumstances in which the potential for reidentifiability is exacerbated, as, for example, in the context of donors affected by rare disorders, due to the small population size, uniqueness of their genotype, or media publicity, which could allow the discovery of personal data linked to the genetic information (Isasi et al., 2012).

Toward a Policy Framework

For scientists, research consortia, bio-repositories, and funding bodies, we envisage a system for data sharing grounded on the principles of good governance that ensures a fair balance between individual interests and public benefits. Such a system should rely on establishing different thresholds for data sharing to minimize the chances of triangulation of a particular data set with other data sets that could further facilitate the reidentification of a participant (Kaye, 2012). These thresholds should be situated along a continuum between overtly identifiable to potentially identifiable data/samples (Rodriguez et al., 2013). They should be subject to ongoing reassessment to reflect the pace of scientific discoveries, consider changing public attitudes, and determine contemporaneous concerns of participants with regards to the meaning of individual privacy and attendant expectations regarding the scope of the fiduciary duties of data/sample custodians.

Moreover, the goal of open science and the principles of transparency, autonomy, and beneficence argue in favor of a system of broad informed consent to sharing genotypic and phenotypic data of hPSC lines, subject to appropriate governance (Knoppers et al., 2011; Isasi et al., 2012). A robust consent process entails empowering participants to make their own risks-benefits assessment before participation. It also requires improving genetic literacy (Knoppers, 2010; Kaye, 2012; Rodriguez et al., 2013). To that end, the consent process should explicitly address data-sharing scenarios and their implications for the protection of participant's privacy and confidentiality. It should further disclose the reasonably foreseeable likelihood of reidentification without overstating the likelihood of these risks materializing, while also acknowledging the nonabsolute effectiveness of available protections.

We propose a framework with criteria for data sharing policies for funding bodies, scientists, research consortia, and biorepositories. Such policy should:

1. Be consistent with participant consent and conform to applicable laws and ethics. Within the consent process the limitations of data protection measures should be disclosed.

2. Establish conditions for releasing data that include a binding, enforceable commitment by researchers and data custodians to not share such data with unauthorized third parties and not to use the data alone or in combination with other data sets to either attempt or create the conditions for the reidentification of an individual participant. To that end, oversight mechanisms should be established.

3. Manage data associated with a given hPSC line (e.g., genomic, epigenomic, phenotypic, and demographic where available) based on a proportional assessment of the risks of individual identifiability, tailored to the nature of cell line derivation (e.g., hESCs versus iPSCs). A cautious approach should be taken when sharing raw sequence reads (e.g., whole genomes and full exomes), short tandem repeats (STRs), SNPs, or other identity profiles, given that they can include sensitive or personal information that is directly identifiable or would facilitate reidentification of otherwise deidentified data. However, research laboratories should be encouraged to share STR profiles of cell lines with bona fide researchers and biorepositories. Identity data (e.g., STRs, SNPs, etc.) should be shared in strict confidence and solely for the purposes of confirming cell line identity for quality control purposes and resolving cases of cell line cross-contamination.

4. Stipulate appropriate sanctions for any breach by those authorized to handle the data.

5. In conformity with recommendation #3 above, make available sensitive or personal data associated with an hPSC line only to bona fide researchers who have provided a protocol that is:

   - Consistent with widely recognized good research practice and with applicable legal and ethical requirements;
   - Aimed at generating new knowledge and understanding using rigorous scientific methods;
   - Intended for publication and sharing of research findings with the scientific community without undue restrictions; and
   - Reviewed by an independent oversight entity.

As the field of hPSC research evolves and with changes in the potential reidentifiability of participants, data stewards should:

1. Make appropriate adjustments to their data sharing arrangements in line with the considerations above; and

2. Avail themselves of research on the concerns of hPSC participants and use such information to guide their data sharing practices.

There are no methods or governance mechanisms that can ensure the absolute protection of participant identity (Rodriguez et al., 2013; Kaye, 2012). Currently, participant reidentification is rare. A proportionate approach to privacy in this context of data sharing should be construed based on reasonably foreseeable risks, thereby distinguishing between perceived and real risks. Such an approach should not rely on worst case or hypothetical scenarios, nor should it relate to situations in which the possibility of identifiability remains negligible (Knoppers, 2010). Most importantly, it should be subject to ongoing reassessment to reflect evolving scientific and IT advances as well as changing public attitudes (which sometimes react to hypothetical scenarios) (Dasgupta et al., 2014). Proportionate criteria for determining what risks are real or which are remote for identifiability are needed to avoid unnecessarily overexpanding privacy regulations that could hinder scientific progress. Moreover, in drafting such criteria, we should question whether in information-rich societies, the goal of complete deidentifiability to avoid privacy-related risks is a realistic or laudable goal (Knoppers, 2010). No amount of legal protection or ethical safeguards can eliminate such risks. Enforceable sanctions (e.g., withholding/terminating actual/future funding or participation in research projects or disclosing misconduct to other funding bodies or stakeholders) against those who misuse data are more realistic and useful legal tools. Furthermore, the use of a more transparent terminology, such as “coded,” that does not refer to “deidentified” cell lines and data—but instead acknowledges the small but potential risk of reidentification—may serve to provide potential participants with a more accurate basis for making informed decisions about whether to assume these risks and permit their cells and data to be used in research.

Acknowledgments

We are grateful to the International Stem Cell Forum and the Stem Cell Network of Canada for funding support. The opinions expressed above are those of the authors alone.

References

Dasgupta I, Bollinger J, Mathews DJ, Neumann NM, Rattani A, Sugarman J. Cell Stem Cell. 2014; 14:9–12. [PubMed: 24388172]

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Science. 2013; 339:321–324. [PubMed: 23329047]

Isasi R, Knoppers BM, Andrews PW, Bredenoord A, Colman A, Hin LE, Hull S, Kim OJ, Lomax G, Morris C, et al. Regen Med. 2012; 7:439–448. [PubMed: 22594334]

Kaye J. Annu Rev Genomics Hum Genet. 2012; 13:415–431. [PubMed: 22404490]

Knoppers BM. EMBO Rep. 2010; 11:416–419. [PubMed: 20448662]

Knoppers BM, Isasi R, Benvenisty N, Kim OJ, Lomax G, Morris C, Murray TH, Lee EH, Perry M, Richardson G, et al. Stem Cell Rev. 2011; 7:482–484. [PubMed: 21279481]

Kreiner T, Irion S. Cell Stem Cell. 2013; 12:513–516. [PubMed: 23642361]

McKernan R, Watt FM. Nat Bio-technol. 2013; 31:875–877.

Rodriguez LL, Brooks LD, Greenberg JH, Green ED. Science. 2013; 339:275–276. [PubMed: 23329035]

Stacey GN, Crook JM, Hei D, Ludwig T. Cell Stem Cell. 2013; 13:385–388. [PubMed: 24094320]