Considerations for Open Release of Genomic Data from Human Cancer Cell Lines

It is the current policy of the National Institutes of Health (NIH) to release comprehensive genomic data from individual research participants in a controlled-access system unless there is explicit consent for open access data sharing to provide some degree of protection to these personal data. Qualified investigators may gain access to controlled data for research purposes following agreement to terms and conditions for responsible use of the data, such as secure storage in computers not linked directly to the internet. This document considers the exceptional circumstances relating to human cancer cell lines commonly used throughout the biomedical research community.

Immortalized cell lines are crucial to biomedical researchers, serving as simplified models of disease states, such as cancer. Genomic sequencing and analysis of approximately 1,000 cancer cell lines are being undertaken by investigators at the Broad Institute, in research partially supported by TCGA, and in collaboration with the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation. The project is called the Cancer Cell Line Encyclopedia (CCLE) and has the important aim to provide this rich resource of data to the research community. The cell lines, available for purchase from the nonprofit American Type Culture Collection (ATCC) or commercial vendors, have become essential reagents in biomedical research. While this statement focuses specifically on CCLE, many of the principles may be applicable to other commercially available human cell lines being used throughout the research community.

Immortalized human cell lines are derived from individual tissue donors so genomic sequence data from these lines could, in theory, be matched to the original donor and his or her immediate blood relatives. The question then arises whether such genomic data sets should be made openly available in NIH-supported databases or, instead, placed within controlled-access systems to provide an extra level of protection for the privacy interests of the donors. Although intended to create a reasonable barrier for legitimate use of data sets in research, the systems for controlled-access release present a hurdle to investigators. For example, an NIH eRA account, which is not available to trainees, is required, and at least one to two weeks pass between request and approval to access data. These procedures necessarily involve certain expectations and work on the part of investigators, and their research organizations, and may discourage some research endeavors.

In considering the balance between protecting the confidentiality of genomic data derived from cell lines and the value of unfettered accessibility (*i.e.*, open access to these data), the following questions were considered.

1. *Does investigation of human cell lines constitute human subjects research per the guidance of the federal Office for Human Research Protections?* Commercially available human cell lines and associated data have been de-identified and no personal information (such as name, address, etc.), is ever made available to researchers who obtain the cell lines. Research investigations involving coded specimens or data are not considered human subjects research per 45 CFR 46.102(f). The Office for Human Research Protections provides further guidance on research involving coded private information or biological specimens. It is quite clear from this guidance that there are no restrictions regarding the sharing of the genomic data from cell lines per the regulations. However, because the NIH Genome-Wide Association Studies (GWAS) policy goes beyond the regulations in its policy expectations, open release of such data would still be a departure from current NIH practices.

2. *Are donors from whom commercially available human cell lines are derived specifically consented for broad release of data?* The informed consent status for donors from whom cell lines originate is unknown by the vendors (and therefore, also unknowable by the investigators analyzing the cell lines). Previously, the distribution of genomic data was not a concern, as it was not possible to link the cell lines or the data generated from them back to the donors. However, the advent of comprehensive genomic analysis introduces the potential of high-density data signatures (e.g., genomic or expression-level data) being matched to an identified secondary data source. The risk to donors is still exceedingly small, as described further in the answer to question 3, but it is strongly recommended that informed consent processes for *prospective* research development of human cell lines (for research or commercial purposes) fully describe and consider any risks associated with broad distribution of genomic data derived from those cell lines.

3. *What is the magnitude of risk to the original donors if the genomic data are released through an open-access as opposed to a controlled-access model?* Individual genomic sequence or variation data can be used to identify an individual or a related family member only if a second identifiable data set exists with which to match the unique variants or, as shown by Gymrek *et. al*[1], if genomic data can be combined with other data sets to make family associations. See also Rodriguez *et al*[2] for further commentary on this issue. The risk for identification to a cell line donor remains remote, although no specific value can be derived for this risk. Similarly, the consequent risks that might be associated with identification of cell line donors remains uncharacterized and likely remote.

4. *Is the amount of genetic variation and genomic sequence data already made public during the course of routine investigations and scientific publication involving these cell lines make further protections of the genomic sequence data pointless?* Genetic variation data on the human cell lines included in the CCLE project are already in the public domain, with SNP array data available from the project web site at the Broad Institute. Similarly, high density

SNP array data are openly available for the well-studied NCI-60 collection of cancer cell lines. Storing genomic sequence data on these cell lines in a controlled-access database would offer no additional protections to the research participants, considering the breadth of data already available.

5. *What are the advantages and disadvantages of open-access as opposed to controlled-access data release for genomic data sets derived from immortalized cell lines?* Although NIH designed the system by which researchers request access to human genomic and phenotypic data to minimize processes (and continues to seek enhancements), it is not instantaneous. Review and approval for access requests generally takes at least one to two weeks. Also, the investigator must have an NIH electronic account to make a request for data access (*i.e.*, the investigator must be designated as a Principal Investigator able to submit grant applications to NIH as confirmed by the investigator's institution). Therefore, a more senior investigator must assume the responsibility for access and analysis on behalf of a student or junior investigator. It is expected that data provided through open-access release will result in more frequent use of the data for potentially a broader range of research purposes by a larger sector of the investigator community.

Conclusions: NHGRI and NCI staff administrating The Cancer Genome Atlas program have considered the risks and the advantages of open release of genomic sequence data from commercially available cancer cell lines, specifically in reference to the Cancer Cell Line Encyclopedia project. The risk of harm to research participants is largely dependent on the risk of identification and is currently thought to be negligible. In circumstances in which genetic data are already available openly, there is no additional risk. Open release of such data sets provides a benefit to biomedical research not fully realized if the data are available only by controlled-access mechanisms. Therefore, it is the position of the TCGA Program that genomic sequence and variation data sets generated from the CCLE will be made available to researchers without requiring Data Use Certification. If the risk calculus for the individuals that donated samples to create these cell lines changes in the future, this policy will be re-examined.

---

[1]Gymrek *et al.*, Science **339**, 321 (2013), Identifying Personal Genomes by Surname Inference.

[2]Rodriguez *et al*., Science **339**, 275 (2013), The Complexities of Genomic Identifiability.