Review

Consent Considerations for Generation and Sharing of Genomic Data

Charles D. Warden *

Posted Date: 25 June 2024

doi: 10.20944/preprints202406.1671.v1

Keywords: data sharing, policy, genomics, genetics

Review

Consent Considerations for Generation and Sharing of Genomic Data

Charles D. Warden †

Independent Researcher, Riverside, CA; cwarden45@gmail.com

† Previously employed as Bioinformatics Specialist in the Integrative Genomics Core at *City of Hope* and will enroll in "*Genetics, Genomics, and Bioinformatics*" PhD Program at *UC-Riverside*.

Abstract: It is important to consider both requirements and limitations for the generation and sharing of genomic data. For example, human genomic sequence data should typically not be deposited publicly without the appropriate "explicit consent" to do so. This includes genetic identifying data that may exist within raw reads for bulk RNA Sequencing (RNA-Seq) data and genomic data from cell lines that are available to purchase. At the same time, data sharing is important for re-analysis and reproducibility/replication in the scientific community. A review of known rules and guidelines for genomic data sharing is provided. This includes specific rules for NIH-funded data for controlled access deposit of HeLa cell line data, including bulk RNA-Seq data. The ability to create "partial" Gene Expression Omnibus (GEO) public deposits with only processed data is described, with search criteria that can identify many "partial" GEO deposits for a variety of data types where reads were not made public for patient privacy concerns. However, it is the opinion of the author that this should be considered a short-term solution, and additional considerations and action should be carried out for genomic data generated in future experiments. Information about how to learn more about what is known for consent for cell line samples is also briefly provided, along with search results for genomic data from widely available cell lines that is deposited in controlled access databases. Finally, it should be made clear that this article presents some amount of opinion. Additionally, open feedback for this preprint is encouraged to further enhance knowledge and communication.

Keywords: data sharing; policy; genomics; genetics

Current Guidelines and Rules

Informed consent is an important point of discussion for generation and sharing of genetic and genomic data [1, 2]. There is detailed information from the United States National Institute of Health (NIH) related to Genomic Data Sharing (GDS, [3]). There are also more general data sharing requirements for NIH-funded studies described by the NIH Data Management and Sharing Policy [4], for grant applications received after January 25th, 2023. One important component is requiring sharing of genomic data for potential re-analysis of NIH-funded projects. Another important component is using *appropriate* data sharing, where some samples may not qualify for certain types of data deposit. Considerations for determining appropriate data deposit include the details of what is described in a consent form for data collection (or re-consented samples) and the date of sample collection. If the necessary consent is not collected, then respecting patient rights is important for limiting certain types of data sharing (even if a project is not funded by the NIH).

There is specific mention of "*cell lines*" in NIH GDS NOT-OD-14-124 [5]. Thus, patient protection for limitations and requirements for genomic data sharing also apply to existing cell lines. While the biospecimen (including the cell line) may be capable of being de-identified, any *genomic data* generated from the samples may contain genetic identifying information or otherwise require treatment as if the data contains biometric information that can be used to identify the patient.

The United States Department of Health and Human Services (HHS) provides decision charts for human subjects requirements [6], which are influenced by the Common Rule (45 CFR 46, [7, 8]). One of the target audiences for these decision charts are Institutional Review Boards (IRBs). There

 doi:10.20944/preprints202406.1671.v1

are currently genomics studies that can receive a "*Not Human Subjects Research*" designation from an IRB [9-11]. However, the precise definition of "*human subject*" used to communicate warnings for genomic data sharing can vary.

For example, the submission page for Gene Expression Omnibus (GEO, [12]) website contains a warning "*WARNING: If you are submitting human data, it is your responsibility to comply with Human Subject Guidelines.*" The linked page for "*Human Subject Guidelines*" describes limitations for deposit of NIH-funded studies as well as non-NIH-funded studies [13], and the linked Frequently Asked Questions section also mentions that data requiring controlled access should be deposited in the database of Genotypes and Phenotypes (dbGaP, [14]). Another option for controlled access deposit is the European Genome-Phenome Archive (EGA, [15]). Either way, data that should not be made publicly available may have been generated from a study designated as "*Not Human Subjects Research*" by an IRB.

Nevertheless, potential genetic identifying information should not be made public without the appropriate consent, and this is a crucial consideration regardless of the source of funding for the experiment.

The Importance of "*Explicit Consent*" for Genomic Data Deposit

The formatting for information related to Institutional Certification for NIH Genomic Data Sharing may change over time. However, upon submission of the initial version of this preprint, there are still instructions consistent with varying consent requirements for samples collected before January 25th, 2015 [16]. In particular, for Extramural researchers, there are separate forms for samples collected with or without consent before January 25th, 2015.

Another important component of the consent is that there be "*explicit consent*" related to the genomic data sharing. This includes specific mention of data sharing and the type of data sharing, such as **controlled access data sharing versus public data sharing**. There cannot be Institutional Certification for samples collected after 1/25/2015 without "explicit consent" related to how genomic data will be shared.

Requirements for NIH-Funded HeLa Genomic Data Deposit

It is known that *HeLa* cells (named after the patient *He*nrietta *La*cks) were not obtained with consent that we currently consider to be appropriate [17, 18]. Accordingly, there was an agreement between the NIH and the Lacks family to deposit HeLa genomic data from NIH-funded studies as **controlled access** [19]. Controlled access deposit for a variety of genomic data generated from the *HeLa* cell line can be found in phs000640.v9.p1: as of 2/20/2024, examples include Whole Genome Sequencing, RNA Sequencing (either bulk or single-cell RNA-Seq), and Hi-C. In the special case of HeLa cells, Institutional Certification for controlled access data deposit is provided by the NIH. The author is not aware of Institutional Certification being provided by the NIH for other cell lines (even if generated before the 2015).

Funding-Independent Precautions for Genomic Data Sharing

The author has obtained raw data and genotypes for himself by completing a HIPAA release form. The Health Insurance Portability and Accountability Act (HIPAA, [20]) provides patient protections. For example, the Health and Human Services (HHS) website makes clear that the HIPAA privacy rule covers genetic data [21], and the HIPAA privacy rule booklet [22] indicates "*[for] purposes of the Privacy Rule, genetic information is considered to be health information.*" Clayton et al. 2019 [23] indicates that genetic information is covered under GINA (Genetic Information Nondiscrimination Act) and this is true "*even if the genetic information is not clinically significant and would not be viewed as health information for other legal purposes*". In the research context, Oza et al. 2023 [24] have a section related to "*HIPAA, PHI, and patient de-identification.*" So, beyond the context of research, there may be additional precautions that are a best practice when storing and sharing genetic data. For example,

 doi:10.20944/preprints202406.1671.v1

avoiding public sharing of genetic identifying information without "explicit consent" may also relate to the patient protections from HHS, beyond NIH requirements for research.

"Partial" GEO Deposits of Only Processed Data

The exact procedure may need to vary for individuals from different organizations. However, in general, a staff member from geo@ncbi.nlm.nih.gov can approve submitting a GEO deposit without raw reads due to the lack of needed consent for public data deposit or the lack of knowledge to confirm the presence of the needed consent for public data deposit. You may be requested by GEO staff to reference the NCBI tracking system identifier in subsequent communications for such "partial" GEO deposits.

The author is not aware of any partial GEO submissions that were considered unacceptable by NCBI Staff. However, journal requirements and/or NIH funding requirements may be different, and part of the goal of this opinion preprint is to offer suggestions of how to avoid the need to make this type of deposit in the future. Also, certain wording by GEO staff may need to be used in a comment for the deposit.

The wording from GEO staff can affect searchability of such "partial" GEO deposits. For example, GEO staff kindly provided search criteria that can identify a number of such deposits ("*("patient privacy") OR ("Raw data not available")*"). As of 5/8/2024, that search yields 1,867 results. GEO search results include both SuperSeries and Series for the same study. Also, importantly, there are separate counts for series and samples, with the count of 1,867 results representing 625 series and 1,242 samples. The number of unique studies is smaller than the count of results (even for the "series"), and more than one sample is typically deposited in a study. So, the exact number for studies and samples is different, but "partial" GEO deposits have been used for a noticeable number of independent studies.

Example of wording for different deposits was also provided by GEO staff, and all of those examples included the wording "*patient privacy concerns*". As of 5/8/2024, a search for using ***"patient privacy concerns"*** as the search criteria identifies 1,734 results (569 "series" and 1,165 "samples"). The data types for such GEO deposits identified with that search criteria include (but are not necessarily limited to) bulk RNA-Seq, single-cell RNA-Seq (scRNA-Seq), Chromatin Immunoprecipitation Sequencing (ChIP-Seq), Illumina Bisulfite Sequencing (BS-Seq), Nanopore Whole Genome DNA Methylation, and Illumina EPIC DNA Methylation Array.

For controlled access deposits in dbGaP, Institutional Certification is required. However, for public data deposits in GEO, there are warnings without this same requirement. Data deposits in GEO can be for either human data or non-human data, where the same consent considerations do not apply to non-human data. Nevertheless, please be aware that you may be releasing genetic identifying data that is not allowed to be made public (regardless of funding organization), even if the deposit is approved by GEO staff. Potentially similar to post-publication review for peer-reviewed publications, such raw data may be need to be removed after an initial acceptance in a public database (possibly even years after the associated publication). So, proactive planning at earlier steps of study design may be a best practice that the author would like to follow.

Checking Consent Information for Commonly Used Cell Lines

All sources of purchased cell lines may not provide the information needed to verify consistent for NIH requirements and/or NIH-based data sharing. So, there may be genomic data sharing limitations that are not immediately obvious to the individuals purchasing the cell lines. This is an important topic, where internal discussions may be taking place.

Nevertheless, American Type Cell Culture (ATCC) has been helpful whenever asking about what additional information is known about the consent for cell lines available to purchase. If there are still limitations in knowledge about consent, then the depositor can consider unknown consent to require treatment as if the sample lacked "explicit consent" for public data deposit of potential genetic identifying information. Labs can contact ATCC with consent questions using Tech@atcc.org.

 doi:10.20944/preprints202406.1671.v1

Controlled-Access Data Deposit for Commonly Used Cell Lines

Examples of controlled access data deposits for cell lines that are available to purchase can be identified through database searches. For example, as of 5/17/2024, a search for "*cell line*" in dbGaP yields >100 results and a similar search for "*cell line*" in EGA yields 100s of results. All of those results have not been carefully checked, and many results are *not* for widely available cell lines. Nevertheless, without requesting controlled access, the following results appear to match the intended goal of the search (at least for some samples): *phs001839.v1.p1*, *phs000825.v1.p1*, *phs003224.v1.p1*, *phs001495.v2.p1*, *phs002903.v1.p1*, *phs000299.v2.p1*, *phs001823.v1.p1*, *phs002202.v1.p1*, *phs001004.v1.p1*, *phs001172.v1.p2*, *phs000938.v1.p1*, and *phs000811.v1.p1*. Related to what was mentioned in an earlier section for HeLa cells, *phs000640.v10.p1* is also among those search results (and also relates to additional specific projects). There may be additional examples of widely available cell lines with precedence for controlled access deposit in dbGaP.

Specific recommendations are provided at the end of the manuscript. However, if a study is NIH funded, then the author recommends checking consent for cell lines ***before*** submitting applications for funding, so that controlled access data deposit can be specified as part of the NIH grant (with approval contingent on funding for a full data deposit in dbGaP). Unfortunately, the author is not aware of a central resource to find consent for established and/or new cell lines.

Guidelines for Generation of New Cell Lines

Genomic data sharing policies are relevant for currently existing cell lines as well as future cell lines. In general, Spector-Bagdady et al. 2019 recommend that "*[consent] should be requested prior to generation*" of cell lines [25]. The author agrees that this is a best practice, even if certain exemptions may be allowed under certain situations. This is also consistent with slides provided to summarize NIH GDS policies [26], which indicate "*For studies using cell lines or clinical specimens created or collected after [January 25th, 2015]…Informed consent for future research use and broad data sharing should have been obtained, even if samples are de-identified*".

There can be additional discussions about whether it may be wise to impose more stringent criteria to samples that can be used to create new cell lines, such as whether there should be "explicit consent" for public data deposit. If so, goals for the frequency of obtaining various levels of informed consent for samples to generate new cell lines can also be discussed. Newly generated cell lines generated with NIH funding should have "*explicit consent*" for genomic data sharing [26], for either public deposit or controlled access data deposit.

Present and Unknown Future Identifiability from Genomic Data

There are a variety of studies that directly or indirectly characterize the identifiability of samples with genetic and/or genomic data [27-29], including some discussions with more of a focus on patients in the medical context [30-33]. For example, Blay et al. 2019 [34] describe the ability to identify individuals through the raw reads for bulk RNA-Seq data. However, *even if a current strategy that can successfully identify the sample is not known*, there are restrictions of sharing sequences from human genomic data.

For example, the Sequence Read Archive (SRA) has a warning that "Human metagenomic studies may contain human sequences and require that the donor provide consent to archive their data in an unprotected database" [35]. A filtering function is provided to reduce the occurrence of human reads deposited publicly through SRA metagenomic data deposits. Especially for protocols like Amplicon-Seq where human reads may only be off-target reads that are still amplified from primers that are not based upon human sequences, a currently existing method that can identify the human subject or patient from that amount of genomic data may not be known. Nevertheless, arguably similar to how caution needs to be taken for patient/subject initials that may match multiple individuals, the author also urges caution in public genomic data sharing.

 doi:10.20944/preprints202406.1671.v1

Recommendations for Improving Privacy-Sensitive Data Deposit

If there is either not consent for public data sharing (which may contain genetic identifying information at the current or future time) or the consent for genomic data sharing is not known, then the author does not consider it appropriate to deposit the raw data publicly. This may be something important to more clearly communicate to the broader community.

At the same time, it is important to be able to be able to re-analyze data starting from raw reads. While partial GEO deposits (only including processed data for bulk RNA-Seq data or scRNA-Seq data) have not created a problem in successful publications so far, the author don't consider this best practice indefinitely.

If it is possible to obtain approval for a controlled access deposit as part of the data sharing plan for an NIH-funded study, then the author recommends that for any data that does not have "explicit consent" for public data deposit.

Acknowledgements

There are several individuals from the local community at my previous institution that provided very helpful advice. However, in the interests of being careful and respectful about policies, those individuals are not acknowledged individually. Instead, I think it is very important to broadly thank everybody that helped improve my understanding with numerous conversations related to genomic data sharing and/or general consent-related questions. I would also like to thank Marcia Miller for general feedback to improve communication in an earlier draft of the preprint. The content and opinions expressed are solely the responsibility of the author and does not represent the views, opinions, policies, or procedures of any other individuals or entities. For example, the author does not wish to give readers the impression that there was complete agreement for all points under discussion with everybody (for all steps in the process of writing this article).

More broadly, I would also like to thank NIH Genomic Data Sharing (GDS) support for helpful discussions, as well as an additional NIH staff member that helped with providing specific feedback related to controlled access HeLa genomic data sharing. I also very much appreciate the assistance from GEO staff to help identify a search function to help systematically identify many "partial" GEO deposits.

I am also very thankful for the assistance provided by ATCC. At least in my experience, support was consistently helpful when asked about details related to cell lines (such as questions related to what is known about consent).

In the interests of open and transparent research with communication of gradually increasing formality, notes for the introductory and general concepts within this manuscript are also available in the following blog post: https://cdwscience.blogspot.com/2020/04/notes-on-limits-for-data-sharing.html. Some of the references for publications related to genetic identifiability are also listed in another blog post: https://cdwscience.blogspot.com/2020/03/testing-limits-of-self-identification.html.

Additional Reading and Resources

In general, there is useful content in "*The Lost Family*" by Libby Copeland (from 2020), which is referenced in Episode #131 of the *DNA Today* podcast. Additionally, "*The Lost Family*" references another book that readers may find helpful: there is a collection of chapters written by various authors and edited by Mark A. Rothstein in "*Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era,*" providing background knowledge in a work that was published in 1997.

There is also content related to learning about genetic markers within the edX course "*Genetics: The Fundamentals*" from Massachusetts Institute of Technology (MIT), which can provide more information related to the brief discussion about genetic identifiability with a limited number of markers. There is also discussion of distinguishing individuals, brief review of calculations/applications for Identity By Descent (IBD) segments, as well other experiences with

 doi:10.20944/preprints202406.1671.v1

genetic testing in the MITx edX course "*Genetics: Population Genetics and Human Traits*". However, those courses also largely cover content less directly related to this discussion.

I also found the Coursera "*Design and Interpretation of Clinical Trials*" course from Johns Hopkins to be useful in learning about some of this material before writing this particular preprint (as described towards the end of this blog post). However, I did not specifically contact anyone related to that course to check an acknowledgment is OK, and I therefore moved mention of that course to this slightly different section.

In general, one goal of the preprint is to potentially find others with similar interests. So, feedback related to additional references or resources is appreciated.

Author Contributions: C.D.W initiated discussions, sought various types of information/education, performed research, and wrote the paper.

Conflict of Interest: The author declares no conflict of interest.

References

1. Zieger, M., Y. Joly, and M.E. D'Amato, *On the ethics of informed consent in genetic data collected before 1997.* Nature, 2024. **627**(8003): p. 271-271.
2. Martyn, M., et al., *Secondary use of genomic data: patients' decisions at point of testing and perspectives to inform international data sharing.* European Journal of Human Genetics, 2024.
3. National Institutes of Health, *Genomic Data Sharing Policy Overview.* https://sharing.nih.gov/genomic-data-sharing-policy/about-genomic-data-sharing/gds-policy-overview, (2/20/2024).
4. National Institutes of Health, *Data Management & Sharing Policy Overview.* https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview, (2/20/2024).
5. National Institutes of Health, *NIH Genomic Data Sharing Policy.* https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html, 2014. **8/27/2014**.
6. Department of Health and Human Services, *Human Subject Regulations Decision Charts: 2018 Requirements.* https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts-2018/index.html, (5/16/2024).
7. Department of Health and Human Services, *Federal Policy for the Protection of Human Subjects ('Common Rule').* https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html, (2/20/2024).
8. Department of Health and Human Services, *45 CFR 46.* https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html, (5/16/2024).
9. NIH National Human Genome Research Institute, *Privacy in Genomics.* https://www.genome.gov/about-genomics/policy-issues/Privacy, (5/16/2024).
10. Gibson, G. and G.P. Copenhaver, *Consent and Internet-Enabled Human Genomics.* PLOS Genetics, 2010. **6**(6): p. e1000965.
11. Lowrance, W.W. and F.S. Collins, *Identifiability in Genomic Research.* Science, 2007. **317**(5838): p. 600-602.
12. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic Acids Research, 2012. **41**(D1): p. D991-D995.
13. National Center for Biotechnology Information, *GEO Frequently Asked Questions: Human Subject Guidelines: Can I submit data derived from human subjects?* https://www.ncbi.nlm.nih.gov/geo/info/faq.html#patient, (2/20/2024).
14. Tryka, K.A., et al., *NCBI's Database of Genotypes and Phenotypes: dbGaP.* Nucleic Acids Research, 2013. **42**(D1): p. D975-D979.
15. Lappalainen, I., et al., *The European Genome-phenome Archive of human data consented for biomedical research.* Nature Genetics, 2015. **47**(7): p. 692-695.
16. National Institutes of Health, *Completing an Institutional Certification Form.* https://sharing.nih.gov/genomic-data-sharing-policy/institutional-certifications/completing-an-institutional-certification-form, (2/20/2024).
17. Beskow, L.M., *Lessons from HeLa Cells: The Ethics and Policy of Biospecimens.* Annual Review of Genomics and Human Genetics, 2016. **17**(1): p. 395-417.
18. Skloot, R., *The immortal life of Henrietta Lacks*. 2017: Broadway Paperbacks.
19. Hudson, K.L. and F.S. Collins, *Family matters.* Nature, 2013. **500**(7461): p. 141-142.
20. Department of Health and Human Services, *Health Information Privacy.* https://www.hhs.gov/hipaa/index.html, (2/20/2024).

 doi:10.20944/preprints202406.1671.v1

21. Department of Health and Human Services, *Does the HIPAA Privacy Rule protect genetic information?* https://www.hhs.gov/hipaa/for-professionals/faq/354/does-hipaa-protect-genetic-information/index.html, (2/20/2024).
22. Department of Health and Human Services, *Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule: Part 7: What Health Information Is Protected by the Privacy Rule?* https://privacyruleandresearch.nih.gov/pr_02.asp.
23. Clayton, E.W., et al., *The law of genetic privacy: applications, implications, and limitations.* Journal of Law and the Biosciences, 2019. **6**(1): p. 1-36.
24. Oza, V.H., et al., *Ten simple rules for using public biological data for your research.* PLOS Computational Biology, 2023. **19**(1): p. e1010749.
25. Spector-Bagdady, K., et al., *Biospecimens, Research Consent, and Distinguishing Cell Line Research.* JAMA Oncology, 2019. **5**(3): p. 406-410.
26. National Institutes of Health, *NIH's Genomic Data Sharing Policy.* https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy_Overview.pdf, (2/20/2024).
27. Russell, D.A., et al., *Developmental validation of the illumina infinium assay using the global screening array (GSA) on the iScan system for use in forensic laboratories.* bioRxiv, 2022: p. 2022.10. 10.511614.
28. Kim, J. and N.A. Rosenberg, *Record-matching of STR profiles with fragmentary genomic SNP data.* European Journal of Human Genetics, 2023. **31**(11): p. 1283-1290.
29. Popli, D., S. Peyrégne, and B.M. Peter, *KIN: A method to infer relatedness from low-coverage ancient DNA.* bioRxiv, 2022: p. 2022.10.21.513172.
30. Oestreich, M., et al., *Privacy considerations for sharing genomics data.* EXCLI journal, 2021. **20**: p. 1243.
31. Sholl, L.M., et al., *Institutional implementation of clinical tumor profiling on an unselected cancer population.* JCI Insight, 2016. **1**(19).
32. McGuire, A.L., et al., *Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider.* Genetics in Medicine, 2008. **10**(7): p. 495-499.
33. Wan, Z., et al., *Sociotechnical safeguards for genomic data privacy.* Nature Reviews Genetics, 2022. **23**(7): p. 429-445.
34. Blay, N., et al., *Assessment of kinship detection using RNA-seq data.* Nucleic Acids Research, 2019. **47**(21): p. e136-e136.
35. National Center for Biotechnology Information, *SRA Submission Quick Start: Metagenomic data.* https://www.ncbi.nlm.nih.gov/sra/docs/submit/#metagenomic-data, (6/7/2024).