Science & Society

Raw Genomic Data:
Storage, Access, and
Sharing

Mahsa Shabani,^{1,*}
Danya Vears,¹ and
Pascal Borry¹

Patients are increasingly being encouraged and supported to access and control their own medical and genomic data. We argue that well-established and transparent raw genomic data retention and returning policies are imperative to enable patients to practice their rights to access and control raw data.

Why Store Raw Genomic Data?

Whole genome sequencing (WGS) and whole exome sequencing (WES) for diagnostic and research purposes generate a large volume of raw data. Previous ethical and legal discussions concerning genomic data management have mainly focused on concerns related to interpretation of the results and the return of both primary and secondary findings from these tests. Yet, to date, issues related to the storage and return of raw data have received less attention, particularly in the clinical setting. This is despite the huge potential for long-term data storage to lead to future data reanalysis and reinterpretation in light of new evidence.

Patients' rights to access raw genomic data partly hinge on the data retention and return policies of the test providers. Such policies concern not only the retention and return of reports summarizing the main results of the genetic investigation, but also primary sequencing data [such as bam (binary alignment) or fastq files (unaligned reads)] that allow both the regeneration of primary results and the

reanalysis of data. As WES and WGS become more embedded in diagnostics, it is important to consider both how current data storage policies align with patients' rights and interests to access raw data, and how these rights can (and should) be managed in the clinical setting.

Data Storage Policies and Practices of Clinical Laboratories

Currently, the rules concerning retention and return of raw genomic data must be inferred from general rules concerning medical data retention and patients' medical records. The lack of specific rules for retention and return of raw genomic data means that laboratories are expected to determine their own internal data storage and return policies in accordance with applicable national and international laws, and to communicate these policies to the patients. Professional communities have issued recommendations in order to guide laboratories to adopt best practices in the storage and return of raw data. For example, the American College of Medical Genetics and Genomics (ACMG) recommends a minimum of 2-years storage of file types that would allow regeneration of the primary results as well as reanalysis with improved analytic pipelines (e.g., bam or fastq files with all reads retained) [1].

Some clinical laboratories described their policies relating to storage of raw data in their consent forms, informing the patients about the availability of DNA sequence data for reanalysis and storage of DNA sequence data for various purposes such as test validation or for research. In contrast, some laboratories only indicate the possibility of future policies to incorporate genetic sequence data to permanent medical records (Table 1).

Patients' electronic medical records are considered a likely place for long-term

storage of raw data; however, some position papers or guidelines, such as a report by the UK PHG Foundation and recommendations by the European Society of Human Genetics (ESHG) on WGS are more skeptical about such storage, due to the data protection concerns, including concerns related to unintended access by insurance companies and employers [2,3].

Concerns related to storage of raw data in patients' medical records have pushed some to look for alternative approaches, such as private storage of data by patients. Arguably, transferring files to the patients and allowing them to store data in private storage may mitigate concerns about unintended access by third parties. A recent example is MIDATA, a nonprofit cooperative that offers patients private storage for a wide range of health and personal data, and allows patients to decide who should have access to the data and for which purposes. Private storage of data could also address logistical concerns, such as infrastructure and the cost of storage that could burden public health care providers. However, the affordability of commercial private storage services for all patients may be of concern, and appear to be in contrast with the values of equity and solidarity in health care. Beyond the financial concerns, private management of storage and access to genomic raw data may require a certain level of awareness and competence among patients about the implications of long-term storage of raw data.

Patients' Access to Raw Data

Patients' access to raw genomic data is considerably influenced by the data retention and returning policies of clinical laboratories. Current consent forms of some clinical laboratories, such as Personalis and Centogene specify that patients (or their physician's on their behalf) may request access to raw data from the

Table 1. Examples from Test Providers' Consent Forms Addressing the Issues Related to Storage and Return of Raw Data

Nationwide Children's Hospital	"... DNA sequence data will also be made available if a re-analysis is ordered by a healthcare provider. The DNA sample, clinical information, and associated DNA sequence data may also be kept in a database at Nationwide Children's Hospital and used for test validation or for research (with privacy assured)." (http://www.childlab.com/Document/Get/239)
Center for Precision Diagnostics	"[...] it is possible that this policy will change in the future in which case your/your child's complete genetic sequence may be incorporated into your/your child's permanent medical records." (http://uwcpdx.org/wp-content/uploads/2016/12/NCGL-GENOME-EXOME-REQ-FORM-and-PROBAND-CONSENT-FORM-DEC2016.pdf)
Personalis	"The raw sequence data generated by the Personalis ACE Clinical Exome test can be requested by the ordering health care provider. That data may become part of the medical record of the person being tested." (http://www.personalis.com/assets/files/Personalis-consent-family-example.pdf)
CENTOGENE	"I give my consent to [...] (7) to provide upon my separate request to me, my physician or – as the case may be – the instructing laboratory the raw data of the genetic analysis." (https://www.centogene.com/fileadmin/pdf/Consents/Consent_CENTOGENE_Consent_form_for_conducting_genetic_analyses.pdf)
Veritas	"I understand that my raw data will not be provided. However, a Variant Call File (VCF) is available upon request and may incur additional fees." (https://www.veritasgenetics.com/documents/veritas-mygenome-informed-consent-form.pdf)
Baylor Miraca Genetics Laboratories	"The FASTQ data files (text-based format for storing nucleotide sequences) from the exome sequencing tests can be requested by an Institution. For FASTQ files to be released by Baylor Miraca Genetics Laboratories to the requesting institution the following information is required for each patient: (1) Institutional Whole Exome/Genome Data Release Form – To be filled out by requesting Institution, and signed by the investigator and patient. The investigator will need to attest that the patient/patient's guardian has been provided with an informed consent document and that this informed consent document specifically authorizes the receipt and analysis of whole exome raw data. (2) Request for and Consent to Release of Information from Individuals' Records Form – to be filled out by requesting institution and patient. Required for every patient." (https://www.bcm.edu/research/medical-genetics-labs/index.cfm?pmid=25544)

clinical laboratory. In contrast, laboratories such as Veritas make it clear that patients cannot access their raw data (Table 1).

Previous studies have demonstrated a strong preference of individuals to receive raw genomic data [4,5]. Notably, many of the individuals indicated that they would seek their own interpretation of the data once they get access to the files [5]. If support in the interpretation of raw data is sought via genetic specialists, there are concerns that this demand will burden a publicly funded health care system. Other individuals may upload their data to online platforms that provide services for interpretation of raw data, such as openSNP, Promethease, GEDMatch, and Genome Mate Pro. The features of such websites vary, ranging from returning health-related results to nonmedical

traits and genealogy. Similar to direct-to-consumer genetic testing offers, concerns have been raised about returning significant medical information to the individuals without the support and counseling of health care professionals [6]. The questions remain about how much support individuals should receive from their healthcare providers in order to interpret the raw data.

Sharing of Raw Data

Furthermore, returning raw data to individuals allows them to control with whom they want to share their data. The option to share data with family members, clinicians or researchers is currently offered by some projects, such as MyGene2. Sharing for advancement of medical research and improving the predictability of genetic testing are mentioned among the underlying reasons for

individuals who uploaded their data on data sharing platforms such as openSNP [7].

In parallel, for-profit approaches for personal genomic data sharing are offered by some commercial services such as Genos and Invitae. Depending on the companies' business models, individuals may be compensated for licensing their genomic data to biotech and pharmaceutical companies, or receive a share of profit when the company sells their data to third parties [8]. Although for-profit sharing of personal genomic data is not a common practice to date, a growing demand from biotech and pharmaceutical companies for large volumes of genomic data may promote such approaches in the future, and draw attentions to the related ethical concerns.

Concluding Remarks

Despite the profound value of long-term storage of raw data for future diagnostic purposes, the privacy concerns associated with storage of sensitive genomic data discourage some test providers from so doing. Transferring raw data into the patients' medical records may result in access by third parties for unintended purposes. Both transfer of data to patients and development of infrastructure for private data storage is suggested as a way to empower patients to control storage of, and access to their raw data. However, this might lead to a reduction of collective efforts to store raw data in medical files, and force individuals to hire private data management services. Further studies are needed to investigate patients' experiences and attitudes regarding private storage of raw data in view of the associated concerns and benefits.

One major advantage of patients' access to their raw data is that they can decide with whom they want to share their data and for which purposes. Previous studies have demonstrated that individuals are willing to share their personal data for research purposes. Data sharing by patients will be beneficial for both clinical and research purposes, and optimize the use of massive raw data that are generated in the clinical setting. Nevertheless,

the ethical and legal grounds of emerging (for-profit and nonprofit) web platforms for personal data sharing should be a subject for ongoing scrutiny by scholars and policymakers.

Regardless, increasing support for patients' access to their raw data would be strengthened by developing transparent data storage and access policies by clinical laboratories. Laboratory policies would play a key role in clarifying the patients' access rights to their raw data, including how and when they can demand it. Patient consent forms is one place that laboratory storage and access policies can be communicated. However, such information is either missing or only partially addressed by the majority of forms in use for genomic sequencing. In addition, if consent to receive raw sequence data is to be truly informed, one might consider that consent forms should also include information about the limitations of self-directed data interpretation services, both from online platforms and health care systems, and the potential privacy concerns related to long-term storage of raw data.

Finally, professional communities could guide laboratories in adopting best practices for the storage and return of raw data, and introduce uniformity to these practices. In so doing, templates for

consent forms should be developed in order to ensure that the necessary information regarding the storage and access to raw data, and the potential for patient choice where appropriate, are incorporated.

Acknowledgements

Mahsa Shabani and Danya Vears are supported by the Research-Foundation Flanders (FWO) Postdoctoral Fellowship.

¹Center for Biomedical Ethics and Law, Department of Public Health and Primary Care, University of Leuven, Belgium

*Correspondence:

mahsa.shabani@kuleuven.be (M. Shabani).

<https://doi.org/10.1016/j.tig.2017.10.004>

References

1. Rehm, H.L. *et al.* (2013) ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* 15, 733–747
2. Van El, C.G. *et al.* (2013) Whole-genome sequencing in health care. *Eur. J. Hum. Genet.* 21, 580–584
3. PHG Foundation (2011) *Next Steps in the Sequence: the Implications of Whole Genome Sequencing for Health in the UK.* http://www.phgfoundation.org/documents/report_next_steps_Sequence.pdf
4. Sanderson, S.C. *et al.* (2017) Psychological and behavioural impact of returning personal results from whole-genome sequencing: the HealthSeq project. *Eur. J. Hum. Genet.* 25, 280–292
5. Middleton, A. *et al.* (2015) Potential research participants support the return of raw sequence data. *J. Med. Genet.* 52, 571–574
6. Badalato, L. *et al.* (2017) Third party interpretation of raw genetic data: an ethical exploration. *Eur. J. Hum. Genet.* 25, 1189–1194
7. Haeusermann, T. *et al.* (2017) Open sharing of genomic data: who does it and why? *PLoS One* 12, e0177158
8. Roberts, J.L. *et al.* (2017) Should you profit from your genome? *Nat. Biotechnol.* 35, 18–20