Genome Medicine

MUSINGS

Our genomes today: time to be clear

Jeantine E Lunshof*[1,2] and Madeleine P Ball[1]

DNA is an identifier. We are not defined by our genome, but our DNA is ours and we can be identified through it. Despite the comments made at the time, it was neither wicked nor tacky when Craig Venter, shortly after the first human genome sequence was published in 2001, publicly revealed that he was one donor of the samples used in Celera's genome sequencing project [1]. Venter later explained that by identifying himself as a donor he had intended to demystify the human genome and to reduce public fears about the potential misuse of genetic information [2].

The old days

Regarding the past puts the issues of identifiability and disclosure of personal and public genomes into perspective. The protection of individuals against the possible negative consequences of disclosure of their genetic information has been a major concern throughout the history of human sequencing. The Human Genome Project (HGP) has placed the protection of individuals at the core of its Ethical, Legal and Social Implications program since its inception as part of the HGP in 1989. In 1983, the US President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, as the designated advisory body to the Congress on these matters, reported on 'Screening and Counseling for Genetic Conditions'. The requirement of confidentiality already ranked first - before 'autonomy' - among the Commission's five recommendations. The Commission recommended that genetics-related information be kept confidential and coded, although, notably, even these nascent recommendations made this conditional: 'whenever that is compatible with the purpose of the data bank' [3].

At the time there existed no doubt that, with the appropriate measures (in particular through coding techniques), anonymity could be preserved. Yet evidence contradicting this viewpoint - in the form of forensic identification of individuals using DNA - already existed in the 1980s. As genome sequencing becomes increasingly widespread and large amounts of data and biospecimens accumulate in many different types of biobanks, addressing the identifiability of individuals is an increasingly pressing issue.

Sharing data, protecting privacy

Biobanks and repositories were established to facilitate the storage and redistribution of samples and data. These efforts seek to meet core scientific requirements of sample and data sharing for the purposes of comparison, re-analysis and avoidance of redundancies in research efforts. Fulfilling the ethical and legal requirement of protecting study participants and, in particular, shielding their identity leads to a fundamental tension between data sharing and privacy.

Researchers, database managers and biobank directors have tried their best to meet both goals by using methods to obfuscate and de-identify biological material and data. These measures are not always successful. In 2007 the National Institutes of Health mandated that genome-wide association studies deposit data in a central database. Aggregate data were thought to be 'safe' and were thus publicly shared by the database of Genotypes and Phenotypes (dbGaP). That policy was immediately modified by dbGaP once it was demonstrated that individual genotypes were identifiable in pooled data [4].

This revision may have been exceptional: regulators are understandably reluctant to admit to incorrect assumptions about data safety. The introduction of restrictions on access to materials that have already been disseminated and have found widespread use does not add to the credibility of regulatory bodies. The recent re-identification of widely disseminated 'de-identified' samples by surname inference led to the removal of some publicly shared information, but not to the removal of associated data and samples from repositories [5]. Notably, the removed data were considered to be compliant with anonymization requirements mandated by the US Health Insurance Portability and Accountability Act, and re-identification occurred despite this. Today it is clear that individuals are identifiable through their unique biological profiles, and evidence of this - for example, from gene expression or microbiome datasets - continues to

*Correspondence: jelunshof@genetics.med.harvard.edu
[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02215, USA
Full list of author information is available at the end of the article

accumulate. Rapid advancements in the genomic sciences have tremendously increased our understanding of human biology. Global collaboration among researchers and sharing of human specimens and data to corroborate findings are key conditions of good scientific practice. Traditional research ethics regarding human subjects is based on different assumptions about scientific practice, assumptions of 'anonymity' being just one example, thereby posing increasing dilemmas to the scientific community.

Personal disclosure

At the other end of the privacy spectrum, opposite to presumed anonymity, is personal disclosure. At the time, many questioned the wisdom of Craig Venter's decision to reveal his inclusion in the compound human genome sequence published by his group [1]. Such a disclosure may have been in conflict with the study protocol, and Venter's action may have warranted more transparency with his colleagues. In addition, with some effort one could eventually extract trait and disease predictions about him and the other DNA donors - some of which could be considered stigmatizing.

However, personal disclosure can have great value. Disclosure has been at the foundation of most patient organizations and research-focused disease interest groups: giving up anonymity and sharing experiences has been for many patients and their relatives the only route to improved diagnostics, treatment and care. The non-profit advocacy organization Genetic Alliance (and its many member organizations) is one excellent example. A case involving mental health information, which is currently considered to be potentially highly stigmatizing, illustrates the great value in disclosure. In 1908, mental health care in the United States was changed forever when Clifford Beers disclosed his personal history of mental illness and the miserable state of care in his autobiography [6], sparking a movement leading to comprehensive mental health reform.

Identifiability today

The prevention of the identification of individuals through meticulous and costly de-identification procedures has kept investigators, statisticians, data managers, ethics committees and oversight bodies busy, as data and sample collections have grown to form large databases and biorepositories.

Yet DNA is an identifier and, as such, all biological material and sequence data can ultimately reveal the identity of their source. While anonymity and confidentiality are promised to study participants, researchers are sharing data by default, as a necessary condition of good scientific practice and as required by funding agencies.

The Personal Genome Project (PGP) has been the first research project to make public sharing of data a reality while avoiding unsustainable promises of anonymity to participants, while their comprehensive genotype and phenotype data are made accessible in the public domain [7-9]. Moreover, the value of the availability of robustly annotated variant sets is increasingly being recognized [10]. Going forward, researchers and participants should consider similar models to the PGP that allow them to build open-access resources; the outcomes and benefits of such clear and open collaborations may well exceed current expectations.

Abbreviations

dbGAP, database of Genotypes and Phenotypes; HGP, Human Genome Project; PGP, Personal Genome Project.

Competing interests

JEL is Ethics Consultant and MPB is Director of Research of the PGP; both serve on a voluntary basis and do not receive a salary or financial compensation from the PGP. The authors declare that they have no other competing interests.

Acknowledgements

JL receives funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013). The funding body had no role in the writing of the manuscript and in the decision to submit the manuscript for publication. The views expressed are entirely the author's own. The authors wish to thank George Church for valuable comments, the staff of the PGP for ongoing inspiring discussion, and the PGP participants for their enthusiasm and contribution to science by putting identified-data sharing into practice. The authors wish to thank Melissa Gymrek and Yaniv Ehrlich for discussion.

Author details

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02215, USA. [2]Section Molecular Cell Physiology, VU University Amsterdam, De Boelelaan 1085, 1081HV Amsterdam, The Netherlands.

Published: 27 June 2013

References

1. Kennedy D: **Not wicked, perhaps, but tacky.** *Science* 2002, **297:**1237.
2. Venter JC: **A part of the human genome sequence.** *Science* 2003, **299:**1183-1184.
3. President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research: *Screening and Counseling for Genetic Conditions: A Report on the Ethical, Social, and Legal Implications of Genetic Screening, Counseling, and Education Programs*. Washington, DC: Government Printing Office; 1983:6.
4. Zerhouni EA, Nabel EG: **Protecting aggregate genomic data.** *Science* 2008, **322:**44.
5. Rodriguez LL, Brooks LD, Greenberg JH, Green ED: **The complexities of genomic identifiability.** *Science* 2013, **339:**275-276.
6. Beers CW: *A Mind That Found Itself*. Project Gutenberg; 1908 [http://www.gutenberg.org/files/11962/11962-h/11962-h.htm]
7. **Personal Genome Project** [**Error! Hyperlink reference not valid.**]
8. Lunshof JE, Chadwick R, Vorhaus DB, Church GM: **From genetic privacy to open consent.** *Nat Rev Genet* 2008, **9:**406-411.
9. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, Angrist M, Bhak J, Bobe J, Callow MJ, Cano C, Chou MF, Chung WK, Douglas SM, Estep PW, Gore A, Hulick P, Labarga A, Lee JH, Lunshof JE, Kim BC, Kim JI, Li Z, Murray MF, Nilsen GB, Peters BA, Raman AM, Rienhoff HY, Robasky K, Wheeler MT, *et al.*: **A public resource facilitating clinical use of genomes.** *Proc Natl Acad Sci U S A* 2012, **109:**11920-11927.
10. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG: **Clinical Genomic Database.** *Proc Natl Acad Sci U S A* 2013. doi: 10.1073/pnas.1302575110.

doi:10.1186/gm456
Cite this article as: Lunshof JE, Ball MP: **Our genomes today: time to be clear.** *Genome Medicine* 2013, **5:**52.