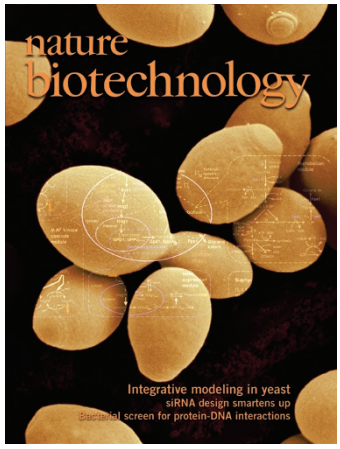
The discombobulation of de-identification

To the Editor:

In 2005 one of us (B.M.K.) coauthored a Correspondence in your pages entitled "The Babel of genetic data terminology," which warned of a dangerously inconsistent and confusing set of terms in the literature describing the identifiability of genetic data[1]. We now are writing to report that, in the intervening decade, the literature has become even more discombobulated with regard to terminology. Here we summarize the Babel-like lexicon for de-identified data and provide our own suggestions for harmonizing terms.



The benefits of next-generation sequencing[2], mobile health apps[3,4], cloud computing[5] and big data analytics have now arrived. They are, however, accompanied by unwelcome friends: namely, a flourishing of novel re-identification techniques that have thrown the idea of guaranteed, total anonymization into question[6–8]. Moreover, international research guidelines are turning away from anonymization for reasons tied to data quality, participant withdrawal and the need to communicate findings and continually link with clinical or other data[9]. Nascent efforts to tie data protection to proportionate and realistic risk assessment are appearing[10,11].

Mandatory policies imposed by funders are pushing researchers toward greatly increased data sharing. Legal duties often require 'de-identification' as a form of privacy protection. Researcher understanding of 'anonymization' often differs in strictness from that which is actually necessary. This almost guarantees over- or under-sharing, which poses risks to participant privacy or research potential, respectively.

Legally, anonymized data is not personal data and thus not subject to personal data protection duties. But there is no consensus definition of anonymization. Although record re-identification codes are sometimes allowed[12,13], law and policymakers tend to define anonymization as "irreversible"[1,14,15]. Occasionally, even indirect identifiers (or quasi-identifiers) seem permissible, as in criteria that ask whether a person's identity "can be readily ascertained"[16]. Still others seem to contradict themselves. The UK Information Commissioner's Office (London), for example, adopts a definition suggesting irreversibleness and conflates anonymized data with pseudonymized data[17] (the latter meaning data that can only be re-identified with access to a deliberately crafted re-identification mechanism). The Global Alliance for Genomics and Health's 2015 Privacy and Security Policy definition labels anonymization as a process that "prevents the identity of an individual from being readily determined by a reasonably foreseeable method"[18]. Later its Data-Sharing Lexicon refined anonymization to mean the "irreversible delinking of identifying information from associated data"[19].

The same holds for other terms describing identifiability. De-identification is often defined as synonymous with irreversible anonymization[1,18,19]. The US Health Insurance Portability and Accountability Act (HIPAA) similarly uses it to refer to data sets to which its anonymization process have been applied. But HIPAA also provides for 'de-identified' data sets to which a re-identification code has been added[20].

Moreover, the term 'anonymous' tends to be used by health researchers to describe information that was collected without direct identifiers, rather than data with identifiers that were later removed[9,15]; however, in recent data privacy instruments such as the European Union's (EU; Brussels) General Data Protection Regulation[14], anonymous is used interchangeably with anonymized, a concept that covers any data that cannot be linked back to an individual.

The need for harmonization of terminology is clear. But what identifiability classification system would best help law and policymakers to regulate de-identification and researchers to understand it?

We believe that 'anonymized data' (or 'anonymous data') should mean data that cannot reasonably foreseeably be re-identified, alone or in combination with other data. 'Pseudonymized data' (often referred to as 'coded data') should mean data that can only be re-identified with access to a deliberately crafted re-identification mechanism. This pseudonymization mechanism can be single- or double-coding, encryption and tokenization, with appropriate safeguards in place. Data that can be re-identified using quasi-identifiers, however, are not pseudonymized. In light of occasional but recurring confusion in the literature on this point, we stress that the mere substitution of direct identifiers with a re-identification mechanism does not result in pseudonymized data, unless it is also shown that its quasi-identifiers do not allow re-identification. Otherwise, the data remain identifiable, and fall within some definitions of 'masked' data[17]. When the data include plain-text direct identifiers, it is 'identified'.

Given the emergence of increasingly sophisticated re-identification attacks[8,21–25], it is now only reasonable to consider genetic data to be anonymized or pseudonymized in narrow circumstances, though we disagree with literature suggesting that anonymization should be abandoned altogether[6,7]. Though even aggregate statistics can allow re-identification of a data set, at some level of generality this ceases to be the case (for example, percentages of US people with a particular single nucleotide variant). The time when the mere removal of direct identifiers was considered defensible anonymization[26] is past. Our dual schema accommodates new techniques, such as secure multiparty computing, homomorphic encryption, *k*-anonymity, and differential privacy[27], without having to explicitly refer to any of them, by making identifiability determinations on a case-by-case, contextual basis[28].

In short, although the details of and difference between techniques to limit identifiability will necessarily be highly significant to technicians applying them to a given data set, our view is that from the perspective of policymakers, the distinctions that are of the highest significance are almost always whether the data have been anonymized or pseudonymized. As to 'de-identification' itself, we believe that the adjective 'de-identified' is ambiguous and confusing to the degree that it should be avoided altogether, whereas the verb 'de-identify' is acceptable to describe any process aiming to limit the identifiability of personal data.

Given the sea of confusion in which the terminology describing identifiability finds

itself[29], none of the terms in the field should currently be used in any text without first clearly defining them. But this precaution merely serves as a stopgap solution to the underlying problems discussed.

Although the simple schema we set out here is compatible with leading law and policy instruments such as the EU Regulation[14] and the Global Alliance's Lexicon[19], it may not yet constitute a consensus approach. We believe that a focus on the de-identification techniques of anonymization and pseudonymization, as defined above, represents the best option. Indeed, new instruments like the EU Regulation seem to be pushing in the same direction, such as by explicitly enshrining pseudonymized data as a distinct category[14], which data privacy law has traditionally declined to do. Broader adoption of these categories would assist law and policymakers in arriving at the most coherent and consistent data-sharing and data-privacy rule sets possible and thereby facilitate researcher and research participant understanding.

ACKNOWLEDGMENTS
The authors acknowledge funding from the Cancer Genome Collaboratory and from Can-SHARE.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Mark Phillips & Bartha M Knoppers

Centre of Genomics and Policy, McGill University, Montreal, Quebec, Canada.
e-mail: mark.phillips2@mcgill.ca

1. Knoppers, B.M. & Saginur, M. *Nat. Biotechnol.* **23**, 925–927 (2005).
2. Solomon, S. in *Clinical Genomics* (eds. Kulkarni, S. & Pfeifer, J.) 403–434 (Elsevier, 2015).
3. Ritter, S. *J. Clin. Trials* **5**, e120 (2015).
4. Zang, J., Dummit, K., Graves, J., Lisker, P. & Sweeney, L. *Technology Science* http://techscience.org/a/2015103001 (2015).
5. Stein, L.D., Knoppers, B.M., Campbell, P., Getz, G. & Korbel, J.O. *Nature* **523**, 149–151 (2015).
6. Ohm, P. *UCLA Law Rev.* **57**, 1701 (2010).
7. Barocas, S. & Nissenbaum, H. in *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (eds. Lane, J., Stodden, V., Bender, S. & Nissenbaum, H.) 44–75 (Cambridge Univ. Press, 2014).
8. Cai, R. *et al. Bioinformatics* **31**, 1701–1707 (2015).
9. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Harmonised Guideline on Genomic Sampling and Management of Genomic Data. Draft Efficacy Guideline E18 (2015)._http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E18_Step2.pdf
10. Article 29 Working Party. Statement on the Role of a Risk-based Approach in Data Protection Legal Frameworks (2014). http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf
11. Cate, F.H., Cullen, P. & Mayer-Schönberger, V. Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines (2014). https://www.oii.ox.ac.uk/archive/downloads/publications/Data_Protection_Principles_for_the_21st_Century.pdf
12. International Standards Organization. Technical Specification 25237. Health Informatics — Pseudonimization (International Standards Organization, 2008).
13. El Emam, K. & Arbuckle, L. *Anonymizing Health Data* (O'Reilly, Sebastopol, California, USA, 2014).
14. Council of the European Union. *General Data Protection Regulation*, first reading, 15039/15 (Brussels, 2015).
15. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Definitions for Genomic Biomarkers, Pharmacogenetics, Genomic Data and Sample Coding Categories. Efficiency Guideline E15 (2007). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E15/Step4/E15_Guideline.pdf
16. *Revised Statutes of Alberta*, as amended (Queen's Printer, 2000).
17. Information Commissioner's Office (UK). Anonymisation: Managing Data Protection Risk, Code of Practice (circa 2012). https://ico.org.uk/media/1061/anonymisation-code.pdf
18. Global Alliance for Genomics and Health. Privacy and Security Policy (2015). https://genomicsandhealth.org/work-products-demonstration-projects/privacy-and-security-policy
19. Global Alliance for Genomics & Health. Data Sharing Lexicon (2015). http://genomicsandhealth.org/work-products-demonstration-projects/data-sharing-lexicon
20. *United States Code of Federal Regulations*. Title 45: Public Welfare, Part 164: Security and Privacy. Office of the Federal Register, National Archives and Records Service, 1996
21. Homer, N. *et al. PLoS Genet.* **4**, 1000167 (2008).
22. Lin, Z., Owen, A.B. & Altman, R.B. *Science* **305**, 183 (2004).
23. Malin, B. & Sweeney, L. *J. Biomed. Inform.* **37**, 179–192 (2004).
24. Schadt, E.E., Woo, S. & Hao, K. *Nat. Genet.* **44**, 603–608 (2012).
25. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. & Erlich, Y. *Science* **339**, 321–324 (2013).
26. Nietfeld, J.J. *EMBO Rep.* **8**, 518 (2007).
27. Article 29 Working Party. Opinion 05/2014 on Anonymisation Techniques (2014). http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
28. Article 29 Working Party. Opinion 04/2007 on the Concept of Personal Data (2007). http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf
29. Institute of Medicine. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research (National Academies, Washington, DC, 2009).