Molecular Systems Biology 8; Article number 612; doi:10.1038/msb.2012.47
Citation: ***Molecular Systems Biology*** **8:612**

www.molecularsystemsbiology.com

molecular systems biology

EDITORIAL

The changing privacy landscape in the era of big data

Molecular Systems Biology **8:** 612; published online 11 September 2012; doi:10.1038/msb.2012.47

Thirty years ago, it was relatively easy to protect one's privacy and remain anonymous. Few computerized systems existed to store our personal information, the internet was so primitive that most were not even aware it existed, and only a few thousand individuals were privileged enough to own a handheld cellular phone. Fast forward to our current day and life—everything has changed. Rapid electronic transactions among individuals and between individuals and entire communities occur on an unprecedented scale, our life stream is continuously digitized and archived—GPS positioning information, cell phone calls, text messages, credit card purchases, e-mails, online social network chatter and even our electronic medical records (Figure 1). In fact, today the marketing department of your neighborhood Target can know before you that your own daughter is pregnant, given changes in purchase patterns (Duhigg, 2012). Long gone are the days of anonymity and privacy.

The life and biomedical sciences have not been shielded from—and are on the verge of massively contributing to—the big data revolution. For example, we recently demonstrated

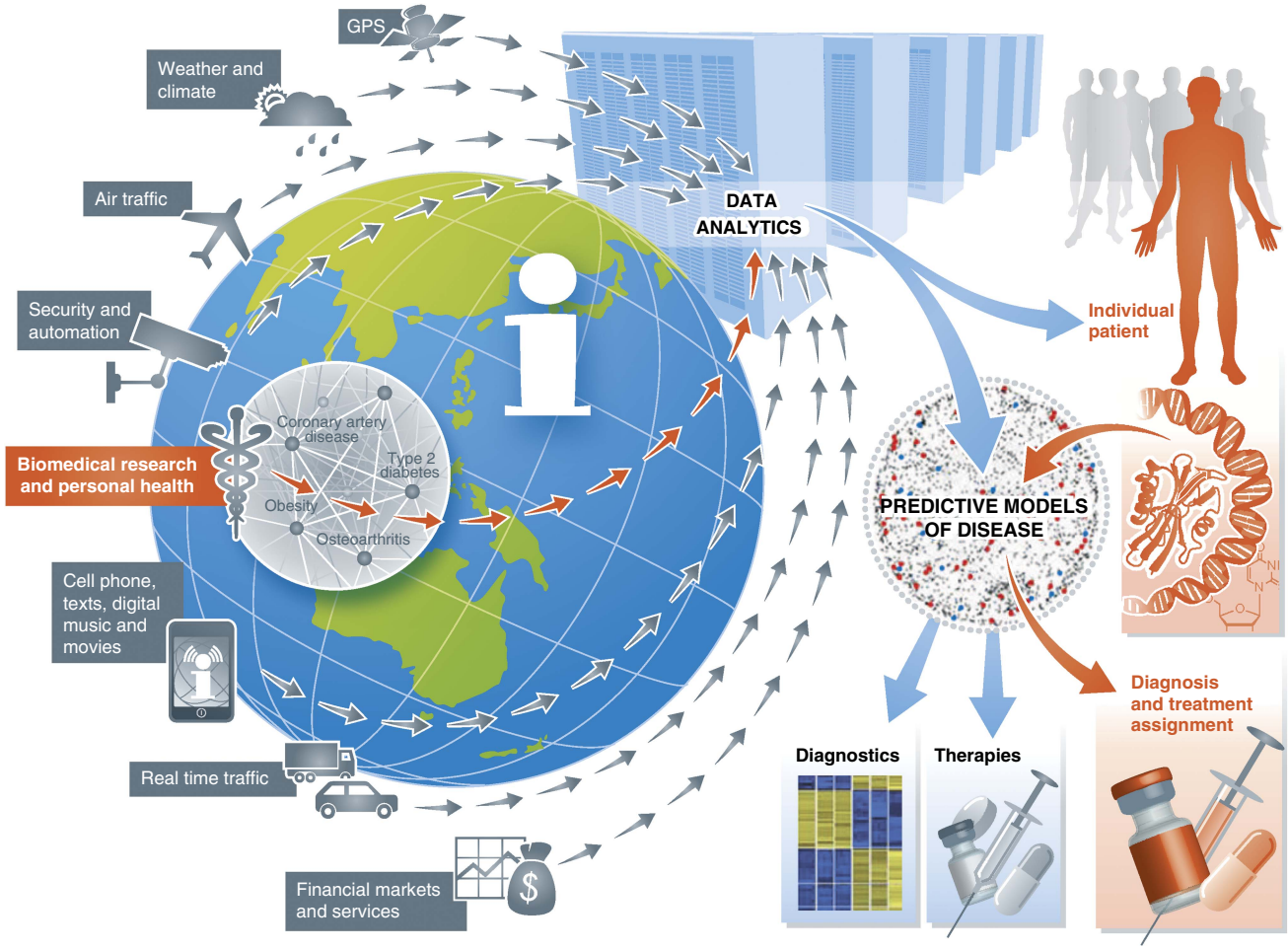

Figure 1 Big data are all around us, enabled by technological advances in micro- and nano-electronics, nano materials, interconnectivity provided by sophisticated telecommunication infrastructure, massive network-attached storage capabilities, and commodity-based high-performance computing infrastructures. The ability to store all credit card transactions, all cell phone traffic, all e-mail traffic, video and images from extensive networks of surveillance devices, satellite and ground sensing data informing on all aspects of the weather and overall climate, and now to generate and store massive data informing on our personal health including whole genome sequencing data and extensive imaging data, is driving a revolution in high-end data analytics to make sense of the big data, drive more accurate descriptive and predictive models that inform decision making on every level, whether identifying the next big security threat or making the best diagnosis and treatment choice for a given patient.

that it is possible to use non-DNA-based information such as RNA abundance measurements to infer a DNA-based barcode that is sufficiently specific to resolve an individual's identity in a collection of hundreds of millions of individual genotypic profiles obtained in a completely different context (Schadt *et al*, 2012). Another study showed that a personal large-scale SNP genotypic profile is sufficient to resolve the participation of one individual to a genome-wide association study, even if the study reports only summary statistics such as allelic frequencies (Homer *et al*, 2008). These examples illustrate that our ability to protect individual privacy in the era of big data has become limited. In particular, the ability to derive DNA-based information from non-DNA-based sources generalizes the issue of data de-identification beyond the area of genotypic data privacy and has thus potentially important consequences for privacy rules in scientific research.

Genomic information has been the main focus of past debates on the protection of privacy and is subject to more legal regulations than other forms of high-dimensional molecular data such as RNA levels. For example, public databases make genome-scale RNA abundance profiles available to anyone with an internet connection. DNA barcodes could in principle be generated from these public data sets, screened against DNA databases kept by government agencies to identify DNA samples associated with, say, an unsolved crime or a terrorist training camp. Government officials could subpoena research records and use the genotypic barcode to identify the matching suspect from the list of study participants.

Expanding laws, locking down relevant databases, and creating greater regulatory burdens to further protect privacy around these types of high-dimensional data represent one set of options. I believe however that such steps would be in vain as the costs of individual molecular profiling technologies continue to fall and as our ability to extract meaning from such data rapidly improves. The life and biomedical sciences are generating information at a furious rate, given the cost of sequencing genomes is dropping at a super Moore's law rate (National Research Council, 2011) and the acquisition of phenotypic data is reaching previously unimaginable scales (Chen *et al*, 2012). Nanopore-based DNA sequencing technologies (Schadt *et al*, 2010) are now on the horizon and will have the potential to generate terabase-scale sequence data in seconds for little or no cost. In 10 years, Google street view cars may not only sample images and Wi-Fi networks, but also sequence everything in their paths and pump in real time that information into big data clouds. It is even not outside the realm of possibilities that just as we derived a genotypic barcode from RNA, so we may be able to derive a barcode from facial parameters by matching facial heritable information—cranial and facial morphological features, morphological features of your ears, skin pigmentation, eye color, iris structure, hair type, hair color—to genomes.

The size of the global digital universe now far exceeds a zetabyte of data (that's 21 zeros or 200 billion 5 GB DVDs), and there is no indication of a slowdown. Data once believed to be harmless in terms of privacy—RNA abundance, cell phone location data—can now be scored in so many dimensions that they can be used to identify an individual. Having reached this tipping point, we now need to understand what information is personally identifiable and to figure out collectively what reasonable expectation of privacy we have regarding such data. What makes the situation even more challenging is that our expectation of keeping information private is a moving target. For example, we usually keep our social security number or medical records private whereas we have no reasonable expectation of privacy for pictures showing our face given we use facial expressions and recognition all the time to communicate in public. Furthermore, our expectations of privacy are rapidly changing. Many individuals today disclose highly personal information on the web and in social networks, loosening our expectations regarding what information should be kept private. Buried within the consents online users click through without ever reading is their explicit approval to allow companies to leverage for whatever purpose they deem appropriate any and all personal information, e-mails, likes and dislikes, political leanings, religious beliefs, and photographs and videos of highly personal scenes in which facial and scene recognition algorithms can be employed to understand your general behaviors and habits, your age, your sex, your friends, types of places you frequent, and the types of products you buy. The same tendency to share all levels of personal details has propagated to the scientific arena as well, with whole genome sequencing and deep molecular profiling carried out now on several scientists who have openly disclosed all data with name attached (Ashley *et al*, 2010; Dewey *et al*, 2011; Chen *et al*, 2012) and direct to consumer genomics companies providing their clients with community 'genome sharing' platforms.

To adapt to this rapidly changing social and technological landscape in ways that serve our individual best interests and that of society more generally, I believe education and legislation aimed less at protecting privacy and more at preventing discrimination will be key. We must inform patients on what is happening in biology and medicine today and explain why high-dimensional data we collect as researchers cannot really be completely de-identified. Direct to consumer genomics companies such as 23andMe may perhaps have a valuable educational role in this context, by enabling anyone to interact with and explore directly their own genomic data, their ancestry, and how others may use their data in the future (from diagnosis to advertising). More and more patients want to share their data with others, to further enable the scientific community to solve problems relating to their condition without being unnecessarily hampered by restrictive rules that prevent, in the name of privacy, a patient from benefiting more directly from data they contribute. Classic consents must therefore transition away from attempting to guarantee individuals' privacy. Rather, new forms of consent should aim at educating research subjects on what the data collected on them can say and the degree to which it can or cannot be protected. Simultaneously, patients should be empowered to have a more vested interest in research outcomes and they should be given more control over sharing their own data with others and with scientists in particular (Box 1).

As big data on individuals becomes more openly accessible, legislative bodies must also be appropriately educated on the consequences of this evolution and expected to enact laws that protect individuals from discrimination based upon their personal information. In the United States, the Genetic Information Nondiscrimination Act and Americans with Disabilities Act

Box 1 The Evolving Informed Consent for Scientific Research

Standard practice for enrolling human subjects in a research study include fully informing potential participants on all aspects of a study including the aims of the study, risks, benefits, costs, and protection of personal privacy. The origins of modern day informed consent for medical research can be traced to the Nuremberg Code in 1947 in an effort to protect participants in research studies (Homan, 1991). However, the omics revolution combined with a far more open data sharing mentality permeating many aspects of society today are driving a new generation of informed consents that put the study participant's ability to openly share data generated on them front and center.

Current Generation Informed Consents

- Single study focused.
- Top-down unidirectional researcher-participant (research subject) relationship.
- Protecting the participant considered among the chief aims.
- Data generation on study participants usually an integral part of the consent.
- Data ownership and terms of use driven by the investigator and/or hosting institution.
- Study participants are counseled to ensure they understand all aspects of the study, although no evidence of understanding is sought or required.
- In most cases, anonymity, privacy, and confidentiality are guaranteed as a key condition for a participant's consent.

'Open Consents' for public resources: The Personal Genome Project Consent (Church, 2005; Lunshof et al, 2008)

Differs from classic informed consent in the following ways:

- Data ownership and terms of use of data no longer driven by study investigator.
- Data are published to the web and made available without restriction.
- Single-study focused, but has broad and open-ended scope (data sharing as an aim).
- Participants agree to reciprocal interaction with researchers.
- Participants must pass an exam to ensure they possess basic genetic literacy, are informed about the public nature of the study, understand the possibility of re-identification, and that some risks are unknown and unpredictable.

Interoperable and Open Consents: The Portable Legal Consent (PLC) (http://weconsent.us/)

Based upon the PGP consent, but altered in the following important ways:

- The PLC can be used across any number of studies.
- If variations of the same PLC form guarantee the same freedoms and creates no more than the same obligations, then it can be certified as interoperable across the PLC network.
- Fully digital, requires no input from a physician or other health/research professional.
- Requires users sign terms of a contract to ensure compliance with data use terms.
- Intended for data already generated, to enable open access of data across many studies.

provide for many such protections, but as patients become more empowered to share their data to achieve greater medical benefit from it, and as we move to more seamlessly map between DNA and more easily acquired high-dimensional phenotypic data to predict with greater ease a greater diversity of human behaviors and disease risks, laws must also evolve to ensure that the rights of patients are protected.

The shift to a more open personal data environment and a greater participation of informed patients will thus need to be accompanied by stricter and broader anti-discrimination regulations. I believe that enforcing such laws will be the condition for our societies to respect individual rights while benefiting from the tremendous potential of big data more openly shared in the life sciences and medicine.

Acknowledgements

I would like to thank Jonathan Wilbanks, Stephen Friend, Jason Bobe, and Daniel Vorhaus for their thoughtful discussion and input regarding research consents.

Conflict of interest

The author declares that he has no conflict of interest.

Eric E Schadt
Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, Institute for Genomics and Multiscale Biology, New York City, NY, USA

References

Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M *et al* (2010) Clinical assessment incorporating a personal genome. *Lancet* **375:** 1525–1535

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D *et al* (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148:** 1293–1307

Church GM (2005) The personal genome project. *Mol Syst Biol* **1:** 0030

Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, Whirl-Carrillo M, Wheeler MT, Dudley JT, Byrnes JK, Cornejo OE, Knowles JW, Woon M, Sangkuhl K, Gong L, Thorn CF, Hebert JM, Capriotti E, David SP, Pavlovic A *et al* (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* **7:** e1002280

Duhigg C (2012) How companies learn your secrets. *New York Times*

Homan R (1991) *The Ethics of Social Research*. London; New York: Longman, ix, 197pp

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4:** e1000167

Lunshof JE, Chadwick R, Vorhaus DB, Church GM (2008) From genetic privacy to open consent. *Nat Rev Genet* **9:** 406–411

National Research Council (US). Committee on a Framework for Developing a New Taxonomy of Disease (2011) *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and A New Taxonomy of Disease*. Washington, DC: National Academies Press, xiii, 128pp

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Hum Mol Genet* **19:** R227–R240

Schadt EE, Woo S, Hao K (2012) Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet* **44:** 603–608