Improving Collective Awareness and Education about the Privacy and Ethical Issues Connected with the Genome Technologies

Lucia Bianchi, Pedro Fernandes, Pietro Lio'
Studio Bianchi (Italy), Instituto Gulbenkian de Ciência (Portugal), The Computer Laboratory, University of Cambridge (United Kingdom)
luciagbianchi@gmail.com, pfern@igc.pt, pl219@cam.ac.uk

Abstract

We may be able to extract knowledge about a person's disease risks from the genome of a cell in his/her body; this knowledge could be extended to some of his/her relatives that share portions of the genome. Current genome technologies may enable insights on personal characteristics, behavior and stress conditions which alter the DNA methylation in different tissues, for example the heart [1]. This paper focuses on the link between health and education, particularly fin what concerns the data that comes from applying genome sequencing technologies [see also 2,3,4]. Let's take a look at the opportunities of accessing good quality education for a little girl in a rural village in an poor African country and those of the son of rich and highly educated parents. The access to a good and significant education in formal educational institutions is sometimes more unevenly distributed than other aspects of life-long learning because opportunities build up on one after another in early school-age years. For example, it is commonly observed that the acceptance in a high quality school creates the basis for continuing the education in similar quality schools. The same pattern is observed in healthcare. in spite the stability of the range in wealth of the countries, and the increase in the individual difference in affluence within each country, particularly in the richest and poorest countries. Scientists in developed countries are setting the basis to a proactive, genome–based, P4 medicine: personalized, predictive, preventive and participatory [5] while, in underdeveloped countries, even basic vaccinations are limited. In a recent intervention, UK Prime Minister Cameron opened up to the possibility that patients could become research patients with their medical details opened up to private and public research. This shift to health crowd-sourcing requires the agreement of boundaries for privacy, data ownership and liability, i.e. requires higher level of collective awareness on genomic benefits and risks. We believe that, as the scarce knowledge about the impact of genome technologies in health is impaired, differences in awareness between poorly and highly educated people could drive both the exploitation of the genetic resources of poor people and the sub-optimality of the medical treatment they will receive or their access to the job market. According to the Genetic Information Nondiscrimination Act, or GINA, an employer cannot fire someone on the basis of not liking something in the employee's genes. The law does not cover life insurance, disability insurance and long-term care insurance. The reason is that if a large number of customers would sign on for policies because they've discovered that they are genetically predisposed to an expensive-to-manage disease, this could bankrupt entire insurance companies. Here we discuss how education programs and online social health platforms facilitate the sharing of information, and new concepts of privacy to enable citizens to make better informed choices and even participating in the decision making process about the services they want to receive, in a democratic participatory fashion, beyond traditional dichotomies, such as cost vs. quality of treatment, economy vs health, demonstrating that it is possible to optimise conflicting choices.

Consent, privacy, crowdsourcing: who owns your genome?

Genome-related technologies, particularly in the areas of synthetic and systems biology will have a huge impact on many aspects of our every day life in health and disease conditions. When it comes to such powerful technologies, the fact that legislation is still navigating slowly in a misty ocean, and the presence of loopholes in the few already existing laws should generate maximum alert. Education becomes even more necessary.

One interesting example is given by the Hela cells which are among the most used experimental models for cancer and cell biology, and were obtained from the tissues of Henrietta Lacks who died of a cancer in 1951. The Hela cells have been an invaluable resource for medical research. Henrietta has never been asked to provide her tissue samples for research. Strikingly such contributions are

usually kept anonymous. A recent study has pointed that anonymity is a false promise in genomic data [6]. The researchers showed they were able to deanonymize publicly available genomic data. They used free genealogical databases that link surnames with genetic markers, called short tandem repeats, on the Y chromosome. Although there is no known biological function associated to these repeats, their characteristics (length in DNA bases and number) are passed from one generation to another (father to son in this case). Then the authors used other pieces of demographic information, such as date and place of birth, which are included in some of the genomic databases and public records to identify 50 donors.

This study suggests that more attention is needed for genetic privacy and the need for better legal protection against genetic discrimination. We found right to say "I own my genome". Should my relatives be asked to express consent, if I decide to disclose my genomic data? Henrietta's relatives claim yes. Kinship is about the fact that we share 1/2 of our genetic material with our mother and 1/2 with our father. We also share 1/2 of our DNA, on average, with our brothers and sisters. The more distant the family relationship, the less genes we have in common. We share a full 1/4 of our DNA with each of our four grandparents, as well as our aunts and uncles. Cousins have 1/8 of their genes in common while second cousins are 1/16 alike. Our genetic likeness continues to drop by 1/2 with each increasingly distant branch in the family tree. In 2007, Rep. Louise Slaughter (D-NY), introduced in the USA Congress the bill Genetic Information Nondiscrimination Act, known as GINA (signed on 21st May 2008). It is noteworthy that the bill was initially proposed in 1995, several years before the first draft of the human genome. Under GINA health insurers cannot deny coverage for genetic causes. Unfortunately several sources have observed that the legislation does not prohibit long-term care insurers, life insurance companies, or disability insurance providers from using genetic data to reject someone seeking a policy. The figure below from Google Trends shows the number of searches for "Genetic information nondiscriminatory act" in the past years. The first peak corresponds to the initial discussion in the Senate and the second peak when the Bill was signed. We observe that the attention has been kept low since then.

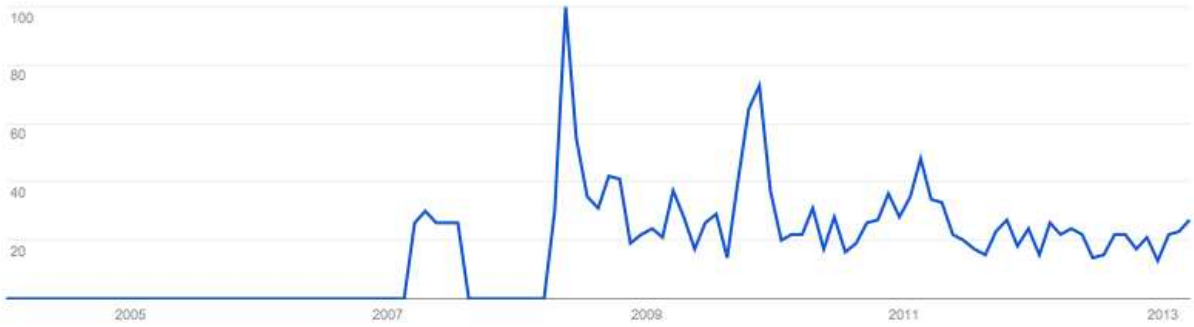

Fig 1: Google trends Web Search Interest: genetic information nondiscrimination act. United States, 2004 - present.

Emerging economic and financial crisis often produce sudden shifts in terms of civil rights. Recently Naomi Wolf has posted a commentary in "The Guardian" on the violent crackdown on the Occupy Wall Street movement. She reported that new documents reveal a coordinated action on the movement not just from FBI and local authorities. The violent action on protesters results from a coordination between FBI and big banks and financial groups [7]. "Reading" in their genome prognosis of deadly diseases, many persons will use life insurances [8]. In turn, during financial crisis the insurance companies would do everything to put in act the genetic testing as precondition for insuring someone.

If I am searching for the cure from a rare and deadly disease, or if I am tracing where traits go in a family can now use crowdfunding to afford the sequencing; this is what Manuel Corpas has recently done at the family level, for 11 individuals [9]. This approach or just the sale of the genomic information, may diffuse in the future; big pharma may become very interested and this could lead to a substantial increase in the number of sequenced human genomes. The idea of using crowdfunding/crowdsourcing to get personal or family genomes sequenced, can also lead to their public disclosure, together with personal health data in portals such as PatientsLikeMe.

PatientsLikeMe is an open sharing personal health platform with about 200,000 patients affected by 1,800 diseases that has started a new type of self-learning healthcare system [10]. Using the level of anonymity of their choice, people connect with others who have the same disease or condition and track and share, in real time, their own experiences, helping other members with comparable characteristics. Researchers, pharmaceutical companies, regulators, providers and nonprofit parties can find grounds to develop more effective products, services and care. The main motivation for sharing personal health data is that they believe that information can change the course of their disease. When personal genomics will be standard procedure, people may share genomic information with friends or completely unknown persons. The privacy worst case scenario will involve two different facts: 1) knowing more about the genetic data of several members of the same family is a step towards being able to make predictions about the genetic data of any descendant; in other words, the larger the number of relatives who get their genome sequenced the more likely each other relative will feel "exposed", i.e. the most likely allelic variants of his/her genome will be known; 2) the combination of predicting personal data and attitudes (say Facebook) and genomic data disclosure will bring public both phenotype and genotype [11].

The complexity of genomic data, embracing multi allelic variants and regulatory circuits (often based on RNA) together with the complexity of a correct patient clinical stratification, often structured in different co-morbidities, would be far from being resolved in the near future. The genomic variants usually point at a huge space of causal models. This condition may induce people affected by a complex disease the will of relaxing on the privacy side in exchange of more insights and understanding on their disease towards a possible cure.

In order to correlate genotype and phenotype we need to develop specific prognostic and predictive biomarkers. The prognostic biomarkers indicate the likely course of the disease in untreated individuals; the predictive biomarkers identify the subpopulations of patients who are most likely to respond to a given therapy. Robust approaches to causal inference are appearing, helping to formulate the causal hypothesis in a non-ambiguous way, determining whether or not the data provides support for that hypothesis. There is an increasing intention to use all the available information; the general population is subdivided into non-overlapping groups (strata), and the information on the proportion of the population in each risk group or condition, is derived from routinely collected statistics or the census.

The multi-parameter evidence synthesis offers a coherent Bayesian analytical framework designed to make rational and exhaustive use of the whole body of information available [12]. It constructs a formal specification of the relationships between data and parameters, which dictates how direct and indirect evidences on the parameters of interest at the genetic, clinical and personal levels can be integrated.

The risk of Facebooking clinical and genetic data

Recently, methodologies such as dimensionality reduction and logistic/ linear regression, have been used to show that Facebook Likes, showing that the technique can be used to automatically and accurately predict a range of individual psychodemographic profiles including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender [11]. People would never reveal any of these traits in public but do not hesitate to provide clues in Facebook because of the soothing and the search for friends approval-related psychological rewarding mechanisms. The predictability of individual attributes from digital records of behavior can easily be applied to large numbers of people without obtaining their individual consent and without them noticing. Commercial companies, governmental institutions, or even Facebook friends could use software to infer attributes such as intelligence, sexual orientation, or political views that an individual may not have intended to share. Importantly, given the ever-increasing amount of digital traces people leave behind, it becomes difficult for individuals to control which of their attributes have been revealed, where and when. Genomic data may quite easily fall in the same pattern.

How to reach the target: education and improved collective awareness

What we need is to set better legal protection against any misuses of data and, most importantly, better awareness through education. We will need better tools in support of the analysis or of the

enforcement of genetic and non genetic privacy; methods for enforcing data privacy, effective anonymity, right to oblivion and non-discrimination. Artificial Intelligence approaches could be devised to discover when genomic data have been frauded. In general, legislation is missing for data mining and privacy policies in social networks related issues, for example context-aware location privacy, particularly in case of detection of unbiased data collection and processing and for enforcing fairness in profiling and targeting. People should be made aware of the possible consequences of this omission, and should acquire a collective awareness of the sensitivity of genomic information. The need to be much better informed on the privacy risks and protections would easily emerge. This would create a natural obligation for public representation in the oversight of genomic data collection.

Collective awareness is a process of collective creation, reinforcing the culture of openness, enabling unprecedented global connectivity of citizens [11]. Collective awareness is different from social innovations, which are generally motivated by and diffused through organizations motivated by profit maximization [13].

The role of education in building collective awareness needs to be assessed at multiple levels.

Formal education is still reflecting the bounded profiles that result from hierarchical nature of their governance. This is an inherited character, transmitted for the sake of keeping control by authority. The general intention is good and the expected result is the continuity of instructional quality as in an inertial system. Learning models of several sorts fit snugly into this rigidness. There is obviously much less room for changes than needed. Learning about the culture of openness in this setting is not really easy. Not impossible, just not easy.

Informal education is a lot more amenable to cultivate openness. Connectivism as a learning model is based on openness, self-organization and adaptation. The models that are being adopted in Open Education Resources (OER) and carried-on into the new massive open online courses (MOOCs) are based on opening access to learning materials first. But if the risk of loosing control is on the way, the management of these education provision resources stays based on the hierarchies, thus limiting the full openness of the provision itself. Some OERs are keen on maintaining the openness everywhere. For the moment, these are the ideal platforms to promote the culture of openness.

Training is, in general, also very amenable to openness. Training needs to be measurable and efficient in transferring skills. In spite of these top level requirements, it is mostly in the hands of the instructors to be flexible while pushing the audiences towards openness. For example, training people using open data resources, open access literature and open source software is in general very possible. It may require some extra effort to build training in that way, but walking along these lines is generally rewarding for instructors and trainees. It is noteworthy that courses to inform lawyers on genomic issues (comprehensive of a wet-lab experience) have been organized in a discontinuous way in several institutions such as Harvard and Bologna.

Formal education is and will be at dire straits with financial issues. Educating less people is not a credible option. Educating up to the population needs has to become affordable, and innovating moves can hardly be justified if they imply irrelevant increases in costs. In other words, proposals that articulate better in this scenario will be the ones that reuse existing resources and attract external funding. Open Education is a very clear way to bring these ingredients, together with the most fundamental components that lead to awareness.

In the particular case of genomics one could argue for a move towards improving sustainability of awareness in the education system, by proposing investment schemes to the sequencing industries. The sequencing industry is producing big data at exponentially growing rates. Revenue is rightfully generated by sequencing “factories” and, downstream, the big pharma, the biotech, the publishing industries generate even more revenue. States that host these industries receive taxes on all that revenue. It would be logical to use part of those tax returns into the education system, precisely to create more awareness.

We propose a way to fairly calculate how much a government should be assigning to this in a way that reflects both the “novelty” of sequencing and the value of the results in the generation of practical knowledge, via the “measure” of valid annotation with a set of ontological terms. This relationship could be of the type Genetic Tax = constant1 x [integration of genomic and phenotype characteristics] – constant2 x [number of sequenced genomes]. Figure 2 shows that the phenotype –genotype integration may depends in a non linear way on the number of genomes sequences and the features identified in the genome. Our capacity to properly address the educational challenge in this field

depends on interdisciplinary and multidisciplinary, social and scientific approaches.

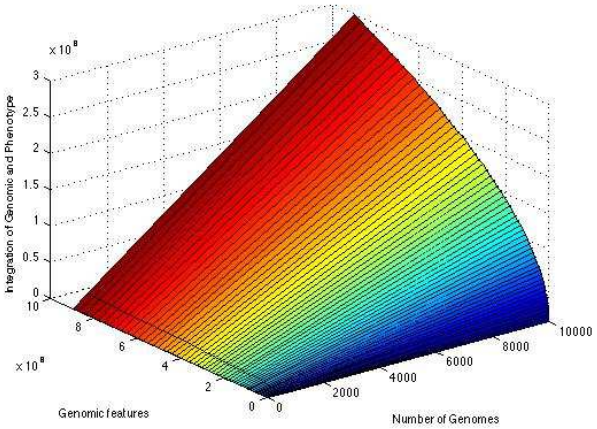

Figure 2: the figure shows, in a semi-quantitative way, in the x-axis the number of genomes, the y-axis the number of features identified in the genome; the z –axis shows the integration of phenotype and genotype information which could follow a non linear relation.

Acknowledgments

PL thanks the FP7 RECOGNITION: Relevance and cognition for self-awareness in a content-centric Internet (257756).

References

[1] Movassagh, M., Choy, M. K., Knowles, D. A., Cordeddu, L., Haider, S., Down, T., Lio' P, Foo, R. S. (2011, November 29). Distinct epigenomic features in end-stage failing human hearts.. Circulation, 124(22), 24

[2] Bianchi, L., & Lio', P. (2009, September 1). La legge e il DNA. Le Scienze, Italian Edition Scientific American, (September 2009 issues). http://www.lescienze.it/

[3] Stajano, F., Bianchi, L., Liò, P., & Korff, D. (2008). Forensic genomics: Kin privacy, driftnets and other open questions. Proceedings of the ACM Conference on Computer and Communications Security, 15-22

[4] Bianchi, L., & Lio', P. (2007). Forensic DNA and bioinformatics. Briefings in Bioinformatics 8(2), 117-128. doi:10.1093/bib/bbm006.

[5] Hood, L., Balling, R., and Auffray, C. (2012). Revolutionizing medicine in the 21st century through systems approaches. Biotechnol. J., 7:1–10.

[6] Gymrek, M, McGuire, A., Golan, D., Halperin, E., Erlich, Y. (2013) Identifying personal genomes by surname inference Science 339, 321-4.

[7] www.guardian.co.uk/commentisfree/2012/dec/29/fbi-coordinatedcrackdown-occupy

[8] Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, et al. (2010) Web-Based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits. PLoS Genet 6(6): e1000993. doi:10.1371/journal.pgen.1000993

[9] Manuel Corposa's blog at http://manuelcorpas.com/

[10] www.Patientslikeme.com

[11] Michal Kosinski, David Stillwell, and Thore Graepel Private traits and attributes are predictable from digital records of human behavior. PNAS March 11, 2013 201218772

[12] Ades AE, Welton NJ, Caldwell D, Price M, Goubar A, Lu G. Multiparameter evidence synthesis in epidemiology and medical decision-making. J Health Serv Res Policy. 2008 Oct;13 Suppl 3:12-22.

[13] F. Sestini (2012) Collective Awareness Platforms: Engines for Sustainability and Ethics. IEEE Technol. Soc. Mag. 31, 54-62