PERSPECTIVES

Personal genomes, participatory genomics and the anonymity-privacy conundrum

VINOD SCARIA*

GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mall Road, New Delhi 110 007, India

[Scaria V. 2014 Personal genomes, participatory genomics and the anonymity-privacy conundrum. *J. Genet.* **93**, 917–920]

Advances in technology have enabled understanding genetic makeup of individuals at a clinical timescale and affordable cost. This has brought about new challenges in the ability to decipher the information content of the genome and be able to act on relevant evidence especially in an environment where the information and evidence is dynamic. The availability of genomic sequences of identifiable individuals in public domain could have far-reaching advantages and open up interesting opportunities, not only to the individual, but also towards understanding the genomic biology. Nevertheless, a framework of social acceptance and regulatory oversight might add to the widespread acceptability of such an approach.

The recent years have seen phenomenal developments in the scale, throughput and consequential unprecedented reduction in the cost of genome sequencing. This change has largely been brought about by an entire gamut of technologies which have enabled miniaturization, large-scale parallelization and improved readout coupled with highly scalable data capture and analysis tools, commonly known as 'next generation' sequencing technologies (Mardis 2008). The cost of whole genome sequencing has been expected to be as low as 1000 US$ in the very near future and is expected to decline further in coming years. This drastic reduction in cost has started to create a dramatic difference in the progress of genome-scale research, evident by the increasing adoption of sequencing techniques to understand genome-scale phenomena. The impact of these technologies in the clinical settings have been discussed to a great extent, and it has been widely speculated and suggested that this would herald the new era of personal genomics and personalized medicine (Brand 2009; Highnam and Mittelman 2012). Genomics tools have also been used increasingly in the recent years to support clinical diagnosis including undiagnosed rare genetic disorders (Maxmen 2011; Jacob *et al.* 2013).

The availability of affordable whole genome sequencing has brought about unprecedented challenges in compilation, analysis, interpretation and translation of high throughput data. One of the major limitations in the application of whole genome sequencing technology in clinical settings has been the lack of understanding the phenotypic correlates of a large number of variations and the paucity of information on the predictability of the phenotype in an individual (Scriver 2004). These limitations primarily arise due to the paucity of tools to assay the enormous spectrum of human phenotypes that constitutes the human phenome and the resulting lack of documentation of phenotypic varieties in all its forms. In fact the human phenome encompasses an enormously large spectrum of traits, many of which are subtle, and sometimes molecular, which adds to the complexity of assaying them. Complete phenome assessment is a tough, if not technically impossible task, given the diversity of the phenotypic space and the costs which need to be incurred to assay dynamic phenotypes over long period of time for large cohorts of people. In addition, it is now well established that epigenetic modifications contribute significantly to the variable expression of a particular phenotype. Nevertheless, the true worth of genomics is in the value of the phenotypic layer of information that could be readily overlaid for integrative analysis (Ghebranious *et al*. 2007; Samuels 2010). This situation offers an interesting conundrum where technological advances have enabled the complete single-nucleotide resolution of whole genome sequencing, but has been severely lagging in the diversity of phenotypes that can be captured for any practical assessment. This conundrum offers two possible alternate solutions, of which one requires significant technological advancements, while the other needs widespread social acceptance. I discuss below each of the possibilities in length.

*E-mail: vinods@igib.in.

Keywords. personal genome; privacy; anonymity; phenome.

The first possible solution is a 'phenome directed genome assessment', whereby genetic variants could be understood based on the phenome of the individual. This entails technological advancements to quantitate a wide spectrum of phenotypes. The availability of ubiquitous and cheap microprocessor electronics-based data quantification technology including activity, sleep, imagery, environment etc. offers a new opportunity to assess phenotypic diversities of individuals of a very large population over a modest timescale. Sensors for physical activity and other vitals including heart rate have been extensively used in clinical settings to aid decision support and patient monitoring. The spectrum of sensors, social networks have been extensively reviewed recently (Swan 2009). The increasing availability of mobile phones capable of modest computing and armed with sensors offer a new opportunity to collect phenotypic diversity of a large population. Increasingly, individuals are making their quantified activities and phenotypes available in public domain, and have emerged into a strong movement popularly named as quantified self, presently spanning many countries and continents across the world (Fleming 2011; Mehta 2011). Nevertheless the number of phenotypic correlates that could be measured realistically over a long period of time still remains limited, without extensive investments in infrastructure and scale-up involving a large number of sensors distributed among a large population of individuals.

The other possibility entails a 'genome directed phenome assessment'. Despite major limitations in the understanding of potential distinct phenotypes that could be predicted from distinct genomic variants at this point in time, this approach may provide an ample opportunity to scale up on already available infrastructure and studies on gene functions revealed from orthologous genes and variants in model systems, without extensive investments in infrastructure and analytic capability. This model could potentially piggy-back on years of understanding and published work on gene functions and potential phenotypes of a very limited number of genes (Smith *et al.* 2007; Thorisson *et al.* 2009). In addition, understanding gene functions through targeted knockouts using a variety of methodologies ranging from transposons, retroviruses (Sivasubbu *et al.* 2007) and very recently TALENS (Miller *et al.* 2011) have become common in academic research and advances in technologies show significant promise in accuracy, scalability and significant reduction in costs.

I argue that a genome directed phenome assessment would be more feasible to achieve as well as be cost-effective. This approach also could significantly reduce the complexity of phenome assessments and also significantly reduce the cost of phenome assessments. The assessment of genome information to direct the phenome though in infancy, has been rapidly progressing in the recent years. The implementation of such a framework would require researchers to revert to the subject as and when new phenotypic information becomes available. Such a possibility does not necessarily preclude privacy as this would only require the availability of electronically traceable but anonymous genome information. Such anonymous but traceable transaction frameworks do exist in other operational areas, e.g. banking and trading. I also argue that 'privacy and anonymity' are not sides of the same coin and could be separated by appropriate technologies borrowed from other areas of day-to-day transactions.

Personal genome information for a number of individuals are already available online. This includes self-revealed personal genomes of individuals like Craig Venter (Levy *et al.* 2007), Jim Watson (Wheeler *et al.* 2008), Stephen Quake (Ashley *et al.* 2010) and others, and a host of anonymous genomes from ethnically and geographically distinct populations and regions like China (Wang *et al.* 2008), Japan (Fujimoto *et al.* 2010), India (Patowary *et al.* 2012), Sri Lanka and Malaysia (Salleh *et al.* 2013). A number of anonymous genomes as part of the 1000 Genome projects and country specific projects like the Korean genome project (http://www.koreangenome.org) and the Singapore Malaysian Project (Wong *et al.* 2013) are also already available in public domain. Apart from this, genotype information at sample level for a large number of individuals who participated in various genotype to phenotype association studies are available as part of the dbGAP (Mailman *et al.* 2007), though many of the datasets are not freely available but are secured under specific license restrictions. The personal genome project (PGP) (Church 2005) is a recent initiative spearheaded to create a publicly available repository of personal genomes for volunteers who would reveal their identity and genome to the public. It has been argued that such an approach with systematic collection of phenotypic correlates would be a valuable tool for research. Apart from such initiatives, individuals who have availed genotyping services through over-the-counter genetic screening services have also made the datasets available in public repositories. Efforts to systematically collect and curate this information has also been initiated, through projects like OpenSNP which has systematically curated both genotype and phenotype information. Availability of such datasets in public domain could have three potential advantages: (i) it would enable identification of genotype–phenotype associations. This has been exemplified by the recent identification of variants associated with subtle traits like striae on the skin (Tung *et al.* 2013). (ii) It would enable the creation of baseline dataset with applications in, e.g. prioritizing rare genetic disease variants by looking at their allele frequencies in the population. (iii) Understanding population structure, admixture and migration at a better resolution. (iv) Analysis for clinically relevant variations, e.g. pharmacogenetic variations to understand and potentially predict population-scale differences in drug efficacy (Giri *et al.* 2014).

Anonymity, privacy and willful revelation of private information have been one of the active areas of debate and research. This has been largely triggered by the rise of social networks and an increasing number of individuals who opt

out for the highly permeable privacy settings offered by many social networks. Increasingly, nonanonymous datasets posted on social media have been used in research in areas as diverse as social sciences, political sciences and recently have been used to track epidemics (Christakis and Fowler 2010). Till date no systematic evaluation of the true value of anonymity with respect to the cost of genome information and insight has been assessed in real-life settings. This would require appropriate availability of information including caveats to whole genome assessment and analysis. The ethical and legal issues of making personal genomes available in public domain have been one of the major challenges in widespread acceptance of this approach. Availability of personal genomes could have potential risks associated with ethical and legal ramifications, e.g. in the area of paternity and ancestry. It could also be potentially mined to predict inherent traits or disease predispositions. Nevertheless many of these risks could possibly be abrogated with appropriate legislation against discrimination (Erwin 2008).

In the recent past, I have been associated with the genomics education website http://www.meragenome.com, which details the technology, benefits and caveats of whole genome sequencing and the extent to which information could be obtained from whole genome sequences. In a recent sample survey conducted online (http://www.meragenome.com/participate), and advertised on social media of prospective participants and genome enthusiasts, we asked whether they would prefer full anonymity at full payment of cost or free availability of whole genome sequencing at no cost. The survey was open to all without restrictions and announced on various personal genome forums. All potential users filled in basic information and were asked to select from four options. The options included: (i) full payment option where the user could maintain anonymity and privacy; (ii) free personal sequencing and payment for analysis where the user had to share the identity stripped data; (iii) free personal genome sequencing and analysis of the user in exchange for willful disclosure of identity and (iv) none of the above. The realistic cost-estimate of 2000$ was fixed as the standard cost of whole genome sequencing. Analysis revealed a surprising trend where majority (over 90%) preferred public sharing of individuality and datasets in exchange for the free availability of genome information. The study is not without caveats though. The major limitations include the small sample size and the unavailability of the complete educational and socio-economic profiles of the sample population. In addition, the sample might also be biased by a specific age-group, and was indeed biased to individuals from India.

The rise of genome enthusiasts who participate willfully in screening tests and share personal information including phenotypic correlates in social media and other networks offer a new opportunity to create a framework towards making use of the rich genetic datasets and phenotypic information for understanding human biology. With widespread adoption of genomic tests and social networks that enable sharing of information, curation and interpretation of genomes and phonemes of all willing participants would thus constitute the next 'big data' challenge, which by far supersedes any big data analysis endeavors mankind has ever embarked upon. I argue that privacy and anonymity are mutually distinguishable and frameworks which could envisage user specified permeability barriers for anonymity and privacy would be the need of the hour. It is also plausible that with adequate safeguards including legislation against discrimination (Erwin 2008), public disclosure of genome and phenomes could become a norm in the future (Lunshof *et al.* 2008; Angrist 2009).

Acknowledgements

The author acknowledges members of the Pan Asian Population Genomics Initiative (PAPGI) and Open Personal Genomics (OpenPGx) consortium for discussions and Dr Abhay Sharma and Dr Sridhar Sivasubbu for critical comments. Funding from Council of Scientific and Industrial Research (CSIR), India through grant BSC0122 (CARDIOMED) is acknowledged.

References

Angrist M. 2009 Eyes wide open: the personal genome project, citizen science and veracity in informed consent. *Per. Med.* **6**, 691–699.

Ashley E. A., Butte A. J., Wheeler M. T., Chen R., Klein T. E., Dewey F. E. *et al.* 2010 Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535.

Brand A. 2009 Integrative genomics, personal-genome tests and personalized healthcare: the future is being built today. *Eur. J. Hum. Genet.* **17**, 977–978.

Christakis N. A. and Fowler J. H. 2010 Social network sensors for early detection of contagious outbreaks. *PLoS One* **5**, e12948.

Church G. M. 2005 The personal genome project. *Mol. Syst. Biol.* **1**, 2005.0030.

Erwin C. 2008 Legal update: living with the genetic information nondiscrimination act. *Genet. Med.* **10**, 869–873.

Fleming N. 2011 Know thyself: the quantified self devotees who live by numbers. The Guardian. (http://www.theguardian.com/science/2011/dec/02/psychology-human-biology) Accessed 10th Nov 2014.

Fujimoto A., Nakagawa H., Hosono N., Nakano K., Abe T., Boroevich K. A. *et al.* 2010 Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat. Genet.* **42**, 931–936.

Ghebranious N., Russell A. W. and Catherine A. M. 2007 Clinical phenome scanning. *Pers. Med.* **4**, 175–182.

Giri A. K., Khan N. M., Basu A., Tandon N., Scaria V., Bharadwaj D. *et al.* 2014 Pharmacogenetic landscape of clopidogrel in north Indians suggest distinct interpopulation differences in allele frequencies. *Pharmacogenomics* **15**, 643–653.

Highnam G. and Mittelman D. 2012 Personal genomes and precision medicine. *Genome Biol.* **13**, 324.

Jacob H. J., Abrams K., Bick D. P., Brodie K., Dimmock D. P., Farrell M. *et al.* 2013 Genomics in clinical practice: lessons from the front lines. *Sci. Transl. Med.* **5**, 194cm5.

Levy S., Sutton G., Ng P. C., Feuk L., Halpern A. L., Walenz B. P. *et al.* 2007 The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254.

Lunshof J. E., Chadwick R., Vorhaus D. B. and Church G. M. 2008 From genetic privacy to open consent. *Nat. Rev. Genet.* **9**, 406–411.

Mailman Matthew D., Michael F., Yumi J., Masato K., Kimberly T., Rinat B., Luning H. *et al.* 2007 The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186. doi: 10.1038/ng1007-1181.

Mardis E. R. 2008 Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402.

Maxmen A. 2011 Exome sequencing deciphers rare diseases. *Cell* **144**, 635–637.

Mehta R. 2011 The self-quantification movement - implications for health care professionals. *SelfCare* **2**, 87–92.

Miller J. C., Tan S., Qiao G., Barlow K. A., Wang J., Xia D. F. *et al.* 2011 A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148.

Patowary A., Purkanti R., Singh M., Chauhan R. K., Bhartiya D., Dwivedi O. P. *et al.* 2012 Systematic analysis and functional annotation of variations in the genome of an Indian individual. *Hum. Mutat.* **33**, 1133–1140.

Salleh M. Z., Teh L. K., Lee L. S., Ismet R. I., Patowary A., Joshi K. *et al.* 2013 Systematic pharmacogenomics analysis of a Malay whole genome: proof of concept for personalized medicine. *PLoS One* **8**, e71554.

Samuels M. E. 2010 Saturation of the human phenome. *Curr. Genomics* **11**, 482–499.

Scriver C. R. 2004 After the genome–the phenome *J. Inherit. Metab. Dis.* **27**, 305–317.

Sivasubbu S., Balciunas D., Amsterdam A. and Ekker S. C. 2007 Insertional mutagenesis strategies in zebrafish. *Genome Biol.* **8** (suppl 1), S9.

Smith D. F., Peacock C. S. and Cruz A. K. 2007 Comparative genomics: from genotype to disease phenotype in the leishmaniases. *Int. J. Parasitol.* **37**, 1173–1186.

Swan M. 2009 Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *Int. J. Environ. Res. Public Health* **6**, 492–525.

Thorisson G. A., Lancaster O., Free R. C., Hastings R. K., Sarmah P., Dash D. *et al.* 2009 HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.* **37**, D797–D802.

Tung J. Y., Kiefer A. K., Mullins M., Francke U. and Eriksson N. 2013 Genome-wide association analysis implicates elastic microfibrils in the development of nonsyndromic striae distensae. *J. Invest. Dermatol.* **133**, 2628–2631.

Wang J., Wang W., Li R., Li Y., Tian G., Goodman L. *et al.* 2008 The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65.

Wheeler D. A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A. *et al.* 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876.

Wong L. P., Ong R. T., Poh W. T., Liu X., Chen P., Li R. *et al.* 2013 Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66.

Received 8 November 2013, in revised form 5 February 2014; accepted 7 March 2014
Published online: 2 December 2014