Published in final edited form as:

Expert Rev Mol Diagn. 2011 June ; 11(5): 457–459. doi:10.1586/erm.11.34.

Privacy and data management in the era of massively parallel next-generation sequencing

Frank S. Ong^{1,2}, Wayne W. Grody^{3,4}, and Joshua L. Deignan³

¹Dept. of Biomedical Sciences Cedars-Sinai Medical Center, Los Angeles, CA 90048

²Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048

³Dept. of Pathology and Laboratory Medicine David Geffen School of Medicine, University of California, Los Angeles, CA 90095

⁴Depts. of Pediatrics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095

Keywords

privacy; patient privacy; data; data management; sequencing; next-generation sequencing; massively parallel next-generation sequencing; full gene sequencing; microarray; oligonucleotide microarray; consent; patient consent; genetics; genetics research; genetic testing; ethics; discrimination; whole-genome; genomics; human genetics; medical genetics; personalized; diagnostic; molecular pathology; diagnostic molecular pathology; bioinformatics; genes; complex

With the creation of the Belmont Report, respect for individuals involved in investigative research was given a high priority, resulting in the ethical standards by which institutional review boards and safety-monitoring policies govern present day research. However, the field of molecular genetics has evolved since that time, starting with the advent of the polymerase chain reaction and dideoxy DNA sequencing and progressing to complex bioinformatics analyses and massively parallel next-generation sequencing [1], where the identity of an individual's digital genetic fingerprint can be traceable. In their editorial, McGuire and Gibbs voiced concern regarding how much information about a particular individual's DNA sequence should be publicly accessible [2]. Compounding this issue is the fact that in order to achieve higher power, genome-wide association studies have been pooling data across international consortia and ethnic populations [3,4]. Even with coded samples and blinded research personnel, issues ranging from subject privacy to discrimination based on an increased disease risk and its impact on any involved third parties undoubtedly predisposes this matter to debate. In addition, there are concerns surrounding the extent of initial consent which should be required and how that may relate to the possible future use of samples for other purposes. The likelihood of unintended discoveries as a by-product of whole-genome analysis obviously complicates matters further. Therefore, as we move into the era of the “thousand dollar genome” [5], the matter of an individual's genomic privacy is of utmost importance. Who has access to the data and how the data are analyzed, stored, and disposed of also need to be closely monitored. As the plethora of genomic data pours out from next-generation sequencing studies and as the annotation, interpretation, and reporting of this genomic fingerprint takes place, ultimately, the fine line between whether this information is confidential, private patient information or whether it belongs in the public domain would have to be clearly delineated.

From the standpoint of an individual, the adequacy of informed consent and the concern about the public dissemination of traceable genetic results, analogous to a genomic digital fingerprint, come to mind. Even though the privacy of subjects is supposedly ensured through the de-identification of samples and the use of various data management models and computer algorithms [6,7,8], additional safeguards will almost certainly be required to go along with the proliferation of databases containing an increasing number of studies and sample sizes, all of them potentially traceable based on the genomic fingerprints. When the privacy risks associated with public data-sharing are discussed with participants as stated in the informed consent (a process which is itself highly variable between different investigators and study centers), subjects who do not want their data publicly broadcast are usually not offered the opportunity to participate in the study, resulting in subject bias and hampering the identification of disease alleles in the population at large. Lin *et al.* proposed masking sensitive data layers from full public view in the tiered data-access approach [9], which has the advantage of minimizing privacy risks, but slows the pace of research and risks discarding potentially useful genomic data. A parallel idea advocated by McGuire and Gibbs is the tiered or stratified consent model where potential study subjects are fully informed about how their genetic data may be transmitted and would have the authority to decide with whom, or if at all, they want their data shared [2]. Most, if not all, clinical laboratories with CLIA certification follow HIPAA recommendations and have standard operating procedures for separating the patient name from the sample upon receipt, barcoding the samples during processing, and encrypting the final test result information to protect patient privacy. However, if genetic tests evolve to encompass large swaths of genomic sequence within the test result, the information could become traceable back to the individual patient.

As sequencing costs continue to fall and the underlying molecular etiologies of polygenic common diseases such as hypertension and diabetes are elucidated, the clinical applicability of massively parallel next-generation sequencing will become feasible. Until that day arrives, Grody and others have suggested that Sanger sequencing of individual genes may be useful for molecular confirmation of atypical cases that are not fully elucidated using limited mutation panels [10], as in the case of *BRCA1/BRCA2*. Even for genes which have been as extensively studied as *BRCA1/BRCA2* (undoubtedly the most intensively sequenced genes in the human genome), previously unseen variants continue to be detected [11]. In today's diagnostic molecular pathology laboratories, the reporting of test results are challenging even for a single common mutation. With full-gene sequencing, and later with partial and full genome sequencing using next-generation technologies, the challenges will be even greater. It is inevitable that an increasing number of sequence variants of uncertain significance (VUS) will cause unnecessary angst for both laboratory directors and clinicians, much as they do now. Furthermore, it is also almost certain that incidental findings, unrelated to the initial indication for testing, will be uncovered. Unlike the case with high-density oligonucleotide microarrays, these incidental findings cannot be avoided by masking, and even if we had the option to mask variants that appear benign or uncertain today, we cannot guarantee that in the future these variants will not be associated with disease [12]. It is imperative that predictive testing using next-generation sequencing be accompanied by adequate and appropriate informed consent, ample counseling and education, maintenance of privacy and confidentiality, and attention to both individual and population-level issues.

From the population perspective, social stigma, both cultural and community-specific, possible affected third party relatives, and international cross-border issues continue to be contested. We are already beginning to see the clinical integration of personal genomic information into the practice of medicine in subtle but gradual ways. Even though the fear may be over-hyped that routine data generation of personal genomic information will lead to

social stigmatization and marginalization, there is very little doubt that increasing a study subject's known risk for a certain disease could have a detrimental impact, whether emotional, social, or financial. We are only beginning to understand the relevance and contribution of genes to complex disease. Even if the correlation of a particular variant is deemed either benign, pathologic, or of unknown significance today, this designation will surely change as more research is conducted to elucidate the function of the gene. Not only does it affect the subject, but given the heritability of these traits, the prognostic health of naive third party relatives may be influenced as well. Beyond the extended family, in an effort to increase sample size and study power, whole ethnic populations across international boundaries are now part of this concern, either knowingly or (more likely) unknowingly. The future use of existing samples and data from subjects who were not initially consented for an alternate study also presents an ethical dilemma for this rapidly evolving field of genomics. Therefore, it is paramount that the data and analysis of massively parallel next-generation sequencing be protected.

Therefore, in this coming age of personalized genomic information, propagated by the rapidly falling cost and turnaround time of sequencing, the cost-to-benefit ratio to both individuals and populations of the information and insight gained from the incorporation of this technology must be carefully weighed. This responsibility falls not only on ethics research policies and institutional review boards, but also on the medical geneticist, the genetics researcher, and even research participants themselves. The American Society of Human Genetics and American College of Medical Genetics have ethical guidelines that practicing geneticists can follow [13,14], but with the rise in popularity of direct-to-consumer marketing of genetic testing, the ethical boundaries can sometimes appear blurred. With increasing cooperation between consortia across various countries, each having their respective ethical standards and cultural practices, this line can be further distorted. In the end, what is at stake is the parity between rich and poor, the informed versus the less-informed, and ultimately our concept of the value of human diversity.

References

1. Rogers YH, Venter JC. Genomics: massively parallel sequencing. *Nature*. 2005; 437:326–327. [PubMed: 16163333]
2. McGuire AL, Gibbs RA. Genetics. No longer de-identified. *Science*. 2006; 312:370–371. [PubMed: 16627725]
3. Kaye J, Boddington P, de Vries J, et al. Ethical implications of the use of whole genome methods in medical research. *Eur. J. Hum. Genet*. 2010; 18:398–403. [PubMed: 19888293]
4. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet*. 2010; 86:6–22. [PubMed: 20074509]
5. Diamandis EP. Next-generation sequencing: a new revolution in molecular diagnostics? *Clin Chem*. 2009; 55:2088–2092. [PubMed: 19855095]
6. Malin BA. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J. Am. Med. Inform. Assoc*. 2005; 12:28–34. [PubMed: 15492030]
7. Sweeney L. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics*. 1997; 25:98–110. [PubMed: 11066504]
8. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. *Trends Biotechnol*. 2003; 21:113–116. [PubMed: 12628367]
9. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science*. 2004; 305:183. [PubMed: 15247459]
10. ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J. Mol. Diagn*. 2008; 10:484–492. [PubMed: 18832462]

11. Easton DF, Deffenbaugh AM, Pruss D, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am. J. Hum. Genet.* 2007; 81:873–883. [PubMed: 17924331]
12. Grody WW. Ethical issues raised by genetic testing with oligonucleotide microarrays. *Mol. Biotechnol.* 2003; 23:127–138. [PubMed: 12632697]
13. Maddalena A, Bale S, et al. Technical standards and guidelines: molecular genetic testing for ultra-rare disorders. *Genet. Med.* 2005; 7:571–583. [PubMed: 16247296]
14. Grody WW, Richards CS. New quality assurance standards for rare disease testing. *Genet Med.* 2008; 10:320–324. [PubMed: 18496029]