Revista de Bioética y Derecho
Perspectivas Bioéticas
www.bioeticayderecho.ub.edu - ISSN 1886-5887

DOSSIER BIG DATA

Big Data in Genomics: Ethical Challenges and Risks

'Big Data' en genómica: retos y riesgos éticos

'Big Data' en genòmica: reptes i riscos ètics

MARC VIA *

OBSERVATORI DE BIOÈTICA I DRET DE LA UNIVERSITAT DE BARCELONA

La Revista de Bioética y Derecho se creó en 2004 a iniciativa del Observatorio de Bioética y Derecho (OBD), con el soporte del Máster en Bioética y Derecho de la Universidad de Barcelona: www.bioeticayderecho.ub.edu/master. En 2016 la revista Perspectivas Bioéticas del Programa de Bioética de la Facultad Latinoamericana de Ciencias Sociales (FLACSO) se ha incorporado a la Revista de Bioética y Derecho.

* Marc Via. Institute of Neurosciences, Universitat de Barcelona. Brainlab, Cognitive Neuroscience Research Group, Department of Clinical Psychology and Psychobiology, Universitat de Barcelona. Institut de Recerca Sant Joan de Déu (IRJSD), Esplugues de Llobregat, Spain. E-mail: mvia@ub.edu.

Abstract

Genomic information is a class of Big Data in expanding use thanks to technological developments. Here, we review three categories of ethical risks and challenges associated with genomic information: privacy issues, the management of incidental findings, and challenges in data storage and sharing. First, we need to implement strong mechanisms to protect privacy, but genomic data faces specific risks and we need to acknowledge the possibility of re-identification. Proper usage of genomic information has to be regulated, including recommendations on incidental finding management. Also, clear policies for data sharing and explicit efforts to promote central repositories of genomic data should be established. However, technology and new applications of genetic information will develop fast and we should anticipate potential new risks.

Keywords: bioethics; genomics; genetics; big data; incidental findings; privacy; data sharing.

Resumen

La información genómica es un tipo de 'Big Data' de uso creciente debido a mejoras tecnológicas. En este trabajo, revisamos tres grupos de retos y riesgos éticos asociados con esta información: riesgos de privacidad, gestión de los hallazgos incidentales y retos en el almacenamiento y compartición de datos. En primer lugar, debemos establecer mecanismos sólidos para proteger la privacidad, pero los datos genómicos presentan riesgos específicos y debemos admitir la posibilidad de reidentificación. Hay que regular el uso adecuado de la información genómica incluyendo recomendaciones para la gestión de los hallazgos incidentales. También hay que establecer políticas claras para compartir datos y fomentar el uso de repositorios de datos genómicos. No obstante, debemos esperar desarrollos rápidos en la tecnología y nuevas aplicaciones de la información genética, y debemos anticiparnos a los futuros riesgos potenciales.

Palabras clave: bioética; genómica; genética; big data; hallazgos incidentales; privacidad; compartición de datos.

Resum

La informació genòmica és un tipus de 'Big Data' d'ús creixent a causa de millores tecnològiques. En aquest treball, revisem tres grups de reptes i riscos ètics associats amb aquesta informació: riscos de privadesa, gestió de les troballes incidentals i reptes en l'emmagatzematge i compartició de dades. En primer lloc, hem d'establir mecanismes sòlids per protegir la privadesa, però les dades genòmiques presenten riscos específics i hem d'admetre la possibilitat de reidentificació. Cal regular l'ús adequat de la informació genòmica incloent-hi recomanacions per a la gestió de les troballes incidentals. També cal establir polítiques clares per compartir dades i fomentar l'ús de repositoris de dades genòmiques. No obstant això, hem d'esperar desenvolupaments ràpids a la tecnologia i noves aplicacions de la informació genètica, i hem d'anticipar-nos als riscos potencials futurs.

Paraules clau: bioètica; genómica; genética; Big Data; troballes incidentals; privadesa; compartició de dades.

1. Introduction

Big Data is a very loosely defined term applied to datasets so massive that are difficult to process using standard methods of database management and statistical analysis. Handling Big Data might be challenging at many levels, but offers opportunities that might be difficult or impossible to obtain from "small data". In a medical context, for instance, we can integrate massive datasets from medical records, sociodemographic surveys, blood and imaging (such as X-rays or ultrasounds) test results and virtually any information that can be cross-linked across datasets. In the era of internet and digitized databases, Big Data analysis is likely to result in great social advances, but it is important to anticipate prospective risks and unintended consequences. In this article, we explore potential ethical risks and challenges associated with the expanding use of genomic information.

Genomic information is Big Data on its own. The suffix -omics is used in biology for fields of study that address the totality of one kind of biological molecule. For instance, proteomics is the field of study that aims at the collective characterization and quantification of all the proteins present in a specific sample. Similarly, genomics analyzes all the variation across the genome of one or more individuals. All the omics, and specifically genomics, are experiencing technical improvements that allow their systematic application in biomedical studies thanks to a dramatic reduction in their cost. A simple example clearly illustrates the reduction in cost of DNA sequencing. After thirteen years of work, the Human Genome Project (HGP) published in 2003 the first complete reference sequence of the human genome at a cost of US$3 billion (US$3,000 million). For over a decade, DNA sequencing costs followed Moore's law, a trend that assumes doubling the power of a technology every two years and characterizes successful technological improvements (Figure 1). From 2007 onwards, DNA sequencing technologies have clearly outpaced Moore's law due to the advent of the so-called 'next-generation' sequencing (NGS) technologies. Thanks to these technological improvements, individual genomes could be sequenced in months at a cost several orders of magnitude cheaper (Wadman, 2008). Today, sequencing a full genome takes just few days —if not hours— and is feasible at an approximate cost of US$1,000.

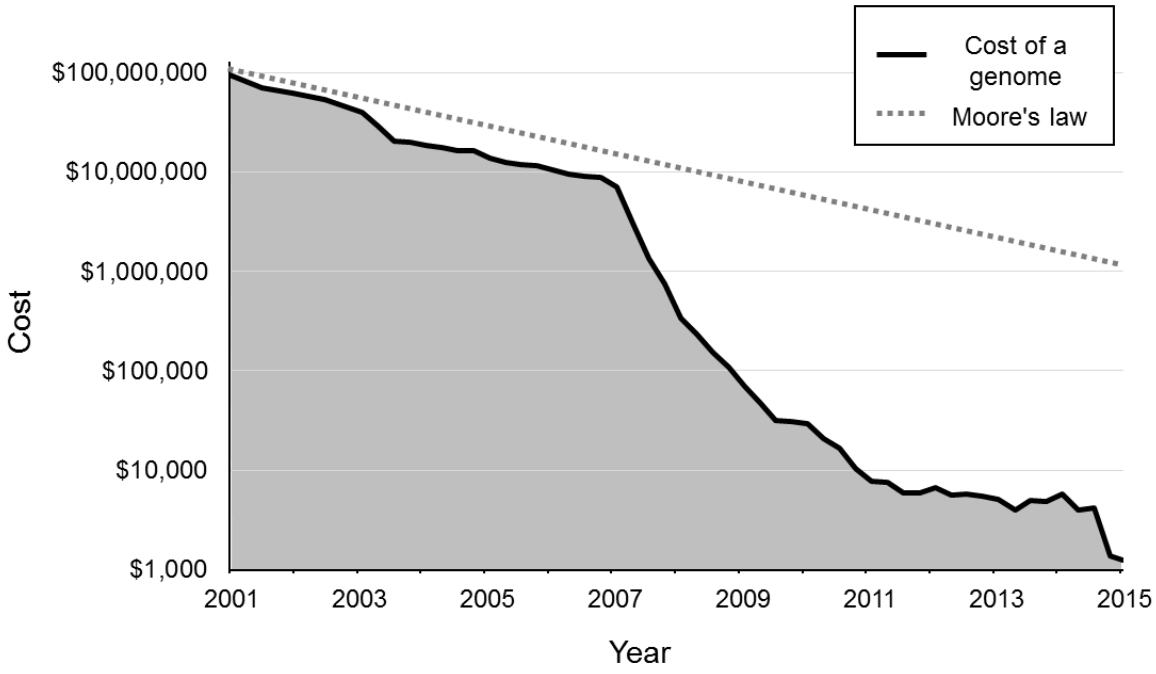

Figure 1. Temporal evolution (2001-2015) of the cost of sequencing a genome. Data estimated by the NIH-NHGRI Genome Sequencing Program and retrieved from Wetterstrand (2017). Note that the Y-axis (cost) uses a logarithmic scale.

The increase in the ability to generate datasets of genomic information has been paralleled with substantial efforts to share and make genomic data available. There is a growing conception that the results arising from publicly funded research projects should be made public and the demands for dissemination, transparency, and responsibility in research also extend to the datasets. In this sense, several resources have been set in place to guarantee access to genomic data. For instance, the 1000 Genomes Project (http://www.internationalgenome.org/) and its predecessor, the International HapMap Project, have made available without restrictions all their data to any user with Internet access. In the US, any federally-funded research that generates genetic data has to be archived in the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap) and a similar repository has been created in the European Union, the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/home). Both dbGaP and EGA have enacted controlled-access mechanisms to protect the privacy of research participants and the confidentiality of their data, while archiving and distributing personally identifiable genetic and phenotypic data resulting from biomedical research projects (Mailman et al., 2007; Lappalainen et al., 2005). Concurrently with these efforts, the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/) has partnered the NIH's Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database

of Japan (DDBJ) to jointly archive raw sequencing data from high-throughput sequencing platforms and make it available to the research community. Moreover, many publishers, especially in Open Access journals, require authors to ensure public access to their datasets before publication.

Despite the benefits that these new genotyping and massive sequencing technologies bring, together with the scientific achievements that allow sharing and public access of genomic datasets, there also exist certain risks and challenges that need to be taken into consideration. Among others, we can group most of these challenges in three main categories: issues associated to privacy, the occurrence of incidental (casual) findings, and challenges associated to the safe management and sharing of genomic data.

2. Privacy Issues

Protecting the privacy of individuals is a classical concern in biomedical research. Data anonymization and informed consent are considered traditional safeguards to protect the privacy of participants in research and in clinical settings (see Chow-White et al (2015) for an excellent review on consent and privacy in the context of clinical and personalized genomics). In addition to them, personal information is usually stored in facilities with restricted access. The transition from locked filing cabinets to digital databases has brought opportunities for big data analysis but also a corresponding set of risks.

A genome is uniquely identifiable but it also has the potential to reveal sensitive information about family members. Since we share 50% of our genome with each one of our parents, siblings, and offspring (and 25% with our grandparents, grandchildren, aunts/uncles...), any privacy breach of genome-based information about the health status of an individual (either present or future) potentially affect other family members. As stated by Chow-White et al. (2015), it constitutes a new category of familial network privacy concern.

In this context, different strategies have been implemented to protect the privacy of participants in genomic research projects. For instance, the Ethical, Legal and Social implications (ELSI) group in the 1000 Genomes Project set up several anonymization practices to preserve privacy, mainly oversampling (i.e. recruiting more individuals than the final number to be included, so not even participants could be sure of their inclusion in the study) and not collecting personal data besides sex. Other projects promote the publication of aggregated data, such as allele frequency or allele-presence information, to simplify data sharing through a web service

and protect participant privacy. One leading example of this approach is the Beacon Project by the Global Alliance for Genomics and Health (GA4GH, http://genomicsandhealth.org/) that provides only allele-presence information. Despite all these efforts, we cannot guarantee total privacy to participants in genomic projects in the era of big data, internet access and data mining algorithms. As pointed out by Chow-White et al., de-identifying and aggregating data is "to hide someone's personal genome signature in a DNA haystack [...]. However, computational data mining algorithms are very good at finding needles in haystacks and linking them to needles in other haystacks" (Chow-White et al., 2015).

A source of potential privacy risks arises from direct-to-consumer (DTC) genetic services. Among them, genetic genealogy services are becoming very popular especially in the USA and they reconstruct genetic ancestry mostly through the determination of uniparentally inherited markers, that is genetic markers that are inherited only through the maternal line (mitochondrial DNA) or through the paternal line (Y chromosome). In addition, other online resources offer free access to databases of genetic genealogy information with search tools to look for potential relatives. To get an idea of the dimension of these databases, Ysearch (www.ysearch.org), one of the largest and most popular genetic genealogy databases, contains more than 190,000 records that include >100,000 different surnames, in the case of patrilineal genealogies. In 2013, Melissa Gymrek and collaborators realized that the information in these databases had the potential to identify anonymous participants in public sequencing projects (Gymrek et al., 2013). They first compared the Y-chromosome markers from three identified public genomes with the information on some genetic genealogy databases, and identified the surname of one of the genomes. Since personal genomes could be identified, the authors then focused on the privacy of current de-identified public datasets. Following a similar approach, the authors could fully identify five anonymous individuals from the CEU collection, multigenerational families of European ancestry in Utah who had been included in several of the most important genetic projects, such as the HapMap and the 1000 Genomes Project. Overall, the privacy of nearly 50 individuals from the CEU pedigrees was breached.

Even aggregated data is not safe from potential re-identifications. For instance, Shringarpure and Bustamante recently showed that it was possible to detect the presence of an individual genome in a dataset that provided only allele-presence information, a so-called "beacon" (Shringarpure and Bustamante, 2015). Since most beacons summarize genomic data of cohorts with a specific disease of interest, disclosing the membership of an individual in a beacon can reveal health-related information about the individual or the individual's family. Although some strategies can be implemented to effectively mitigate privacy risks (Raisaro et al., 2017), zero risk does not exist not even in aggregated datasets.

Although strict legislation to protect privacy is a common practice in most countries, very few of them have enacted comprehensive policies to regulate genomic information. In the USA, the Genetic Information Nondiscrimination Act (GINA) specifically prohibits discrimination based on genetic information in employment and health insurance and some states have extended protection to other forms of discrimination. For instance, California promoted CalGINA to protect individuals from genetic discrimination in housing and education, among others. In the EU the situation is not homogenous regarding genetic information and may vary from country to country. However, the EU Data Protection Directive regulates protection of all sorts of data, including health related information.

3. Incidental Findings

During the course of biomedical investigation, researchers may encounter unexpected findings of potential clinical relevance. For instance, in neuroimaging studies it is not uncommon to detect clinically relevant findings out of the scope of research that pose a challenge to clinicians (Leung, 2013). These are called incidental findings (IF) and several clinical protocols have been set up to guide clinicians on how to handle them. However, with the advent of whole-genome sequencing (WGS) techniques the magnitude of IFs in genomics has vastly increased. According to recent genomic research, the average genome of a typical healthy adult subject contains thousands of genetic variants associated with complex diseases and 25-30 variants associated with rare diseases (1000 Genomes Project Consortium, 2015). In the context of WGS it has been pointed out that IFs should no longer be considered unexpected, but rather anticipated secondary findings regardless of the purpose of the study (Wright et al., 2015).

However, there is still much debate on how to manage IFs in genomic studies. Many questions arise such as who should be responsible to detect IFs and interpret them, which kinds of IFs should be shared with subjects, or how to communicate them. In the clinical setting, there are recommendations on how to report IFs in WGS studies. For instance, the American College of Medical Genetics and Genomics (ACMG) published their recommendations in 2013 including a short list of genes that should be systematically screened in in clinical exome and genome sequencing analyses (Green et al., 2013). After three years of experience in the implementation of these recommendations, they recently published an "updated secondary findings minimum list" with fifty-nine recommended genes (Kalia et al., 2017). In the EU, there is not such a consensus and heterogeneity in the management of IFs is found across different European member states.

However, some coordinated efforts lead by the European Society for Human Genetics (ESHG) are heading in the same direction (Hehir-Kwa et al., 2015).

Regardless of these open questions, it is clear in the clinic that IFs have to be handled in the best interest of the patient. It is not clear, however, how to handle IFs in research participants. In the absence of clinical symptoms or a family history of a specific disease, exploring a person's genome looking for potentially pathogenic variants will likely lead to false positive results and over-diagnosis. Most genetic variants categorized as pathogenic have been characterized only in affected individuals and their relatives, and little is known about their clinical relevance in the asymptomatic general population (Wright et al., 2013).

4. Data Management and Sharing

Another category of challenges when working with genetic data at a genomic scale comes from the amount of data generated. The own nature of this kind of data makes it big and, similarly to other Big Data, it demands facilities to be stored in a secured way and requires tools for an efficient access, analysis and sharing. In a very interesting study, Stephens and cols. have compared genomics with three other major generators of Big Data: astronomy, YouTube and Twitter (Stephens et al., 2015). Projections of the computational needs to the year 2025 showed that genomics will be the most demanding (or on par with the most demanding) in data acquisition, storage, distribution, and analysis.

First of all, there are increasing needs for storage space. Hundreds of thousands of individual human genomes are already stored at the SRA (www.ncbi.nlm.nih.gov/sra/) and twenty of the largest sequencing centers are already using more than 100 petabytes (i.e. more than 102,400 terabytes) of data storage space (Stephens et al. 2015). Storing and sharing genomic data also faces the same kind of potential hazards than any other information in a networked server: transfer speed, power outages, servers crashing, loss of data, or hacking attempts, among others. Problems associated to transfer speed are especially relevant in data of this magnitude. Distribution over the web from central repositories of datasets that can be terabytes in size can be very slow and prone to crashing. In a local context, it is faster sometimes to transfer the data on a physical external drive from one building to another than transferring the data over the net.

Several strategies are in progress to reduce the size of data generated in WGS studies such as improvements in data compression and in sequencing accuracy. For instance, some file formats store only the list of variants relative to a reference sequence (since ~99% of the genome is shared

among all individuals) in what is called as "delta encoding" (Christley et al., 2009). One of the most popular format of delta encoding for genomic information is the Variant Call Format (VCF) and its binary counterpart BCF, initially developed by the 1000 Genomes Project and currently maintained by the Global Alliance for Genomics and Health (GA4GH, http://genomicsandhealth.org/). Reducing the size of data would potentially facilitate storing and sharing. However, these approaches will only slow down the pace of growth of computational needs at best.

An alternative that is becoming increasingly popular is the use of cloud computing services. These services solve (at least part of) the problems in storage capacity, bandwidth and computational power (Chow-White et al., 2015) and have been adopted by major public projects such as the 1000 Genomes Project, that stablished a collaboration with the cloud computing services from Amazon (NIH, 2012). Since many of the analyses on genomic data stored in cloud services can be run remotely, computing resources are optimized. However, these advances also pose different ethical challenges. On the legal side, the cloud server and the genomic data stored in it may be physically located in a different geographic location (i.e. under a different regulatory framework) than the location of the actual research. Moreover, additional security efforts have to be set in place to ensure individual privacy including, but not limited to, new authentication and encryption methods. These are both common problems when using third-party data management services. In this sense, a Framework for Responsible Sharing of Genomic and Health-Related Data has been established by the GA4GH (https://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data).

5. Concluding remarks

We have seen that the use of genomic data faces several ethical challenges, many of which are shared with health-related information and other Big Data sets, but others are specific to the own nature of our genome. First of all, we need to make extra efforts to protect individual privacy, but in the genomic era complete protection may become unfeasible. Every genome is unique but shared with a vast network of relatives. This characteristic makes genomic information prone to specific privacy breaches absent in other generators of Big Data. In this sense, we need to instruct clinical patients and research participants about the risks and benefits of genomic studies. Among other risks, it is important in the consent process to acknowledge the possibility of re-identification. However, we also should educate participants about the importance of sample donation as a pillar of scientific (including medical) progress (Gymrek et al., 2013). A great

example is illustrated by the 1000 Genomes Project in their consent template: "there may be new ways of linking information back to you that we cannot foresee now. [...] We believe that the benefits of learning more about human genetic variation and how it relates to health and disease outweigh the current and potential future risks, but this is something that you must judge for yourself" (International Genome Sample Resource, 2017).

From the regulatory point of view, we need to establish clear policies for data sharing and proper usage of genomic information. Among others, we need to establish recommendations on how to handle secondary findings in genomic studies. These recommendations will facilitate the operations that clinical and research laboratories run on a daily base by providing them with a clear guidance instead of leaving IF management to individual decisions. At the same time, we have to be aware that technological advances and new applications of genetic information will develop fast and beyond our predictions. As stated before, use of health-related data "far outpaces the governance and due diligence of the ethical considerations that need to be addressed" (Bourne, 2015).

In addition to legislative efforts, official agencies also have a pivotal role in the promotion of central repositories of genomic data. These public repositories are essential for the advance of the field by efficiently storing and sharing genomic information while ensuring privacy protection and data confidentiality. In a similar manner as to the prevention of new uses of these data, we should also anticipate potential ethical risks in the storage and sharing of genomic data. Nevertheless, security hazards associated to genomic data might be lower than in other sources of biomedical and health-related information and that might let us buy some extra time to implement preventive measures. In a digital world where information has a value, genomic data is still less tempting than other datasets to attract substantial amounts of interest apart from research and clinical scenarios.

Finally, funding agencies should also be aware of the new paradigm in genomic sciences (and in other omics). Researchers and clinicians now have increasing computational needs that are not adequately addressed in many funding programs. It is very frustrating when research grants implement severe restrictions to cover basic hardware needs, to hire qualified IT staff or even to purchase external hard drive devices.

Genomics is a fast growing field with broad applications and further ethical risks need to be taken into consideration. Among others, the commercialization of genetic products is flourishing especially in the form of direct to consumer (DTC) genomic services. Although DTC services are outside the scope of this article (see, for instance, Chow-White et al., 2015), we anticipate an

increase in the ethical challenges emerging from these products that will parallel the expected advances in personalized medicine.

References

1. 1000 GENOMES PROJECT CONSORTIUM, AUTON A., BROOKS L.D., DURBIN R.M., GARRISON E.P., KANG H.M., KORBEL, J.O., MARCHINI J.L., MC CARTHY, SM, MC VEAN G.A., ABECASIS G.R. "A global reference for human genetic variation". *Nature* 2015; 526(7571):68-74.
2. BOURNE, P.E. "Confronting the ethical challenges of big data in public health". *PLoS Computational Biology* 2015 Feb 9;11(2):e1004073. doi: 10.1371/journal.pcbi.1004073.
3. CHOW-WHITE, P.A., MAC AULAY, M., CHARTERS, A., CHOW, P. "From the bench to the bedside in the big data age: ethics and practices of consent and privacy for clinical genomics and personalized medicine". *Ethics and Information Technology* 2015;17(3):189-200.
4. CHRISTLEY, S., LU, Y., LI, C., XIE, X. "Human genomes as email attachments". *Bioinformatics* 2009 Jan 15;25(2):274-5.
5. *Genetic Information Nondiscrimination Act*, Public Law 110-233, 122 Stat. 881.
6. GREEN, R.C., BERG, J.S., GRODY, W.W., KALIA, S.S., KORF, B.R., MARTIN, C.L., MCGUIRE, A.L., NUSSBAUM, R.L., O'DANIEL, J.M., ORMOND, K.E., REHM, H.L., WATSON, M.S., WILLIAMS, M.S., BIESECKER, L.G.; "American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing". *Genetics in Medicine* 2013;15(7):565-74.
7. GYMREK, M., MCGUIRE, A.L., GOLAN, D., HALPERIN, E., ERLICH, Y. "Identifying personal genomes by surname inference". *Science* 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566.
8. HEHIR-KWA, J.Y., CLAUSTRES M, HASTINGS RJ, VAN RAVENSWAAIJ-ARTS, C., CHRISTENHUSZ, G., GENUARDI, M., MELEGH, B., CAMBON-THOMSEN, A., PATSALIS, P., VERMEESCH, J., CORNEL, M.C., SEARLE, B., PALOTIE, A., CAPOLUONGO, E., PETERLIN, B., ESTIVILL, X., ROBINSON, P.N. "Towards a European consensus for reporting incidental findings during clinical NGS testing". *European Journal of Human Genetics* 2015;23(12):1601-6.
9. LAPPALAINEN, I., ALMEIDA-KING, J., KUMANDURI, V., SENF, A., SPALDING, J.D., UR-REHMAN, S., SAUNDERS, G., KANDASAMY, J., CACCAMO, M., LEINONEN, R., VAUGHAN, B.,

LAURENT, T., ROWLAND, F., MARIN-GARCIA, P., BARKER, J., JOKINEN, P., CARREÑO TORRES, A., RAMBLA DE ARGILA, J., MARTINEZ LLOBET, O., MEDINA, I., SITGES PUY, M., ALBERICH, M., DE LA TORRE,S., NAVARRO, A., PASCHALL, J. & FLICEK, P. "The European Genome-phenome Archive of human data consented for biomedical research". *Nature Genetics* 47, 692–695 (2015) doi:10.1038/ng.3312.

10. INTERNATIONAL GENOME SAMPLE RESOURCE (2017) 1000 GENOMES PROJECT: DEVELOPING A RESEARCH RESOURCE FOR STUDIES OF HUMAN GENETIC VARIATION. "Consent to Participate" http://www.internationalgenome.org/sites/1000genomes.org/files/docs/Informed%20Consent%20Form%20Template.pdf. Accessed April 20, 2017.

11. KALIA, S.S., ADELMAN, K., BALE, S.J., CHUNG, W.K., ENG, C., EVANS, J.P., HERMAN, G.E., HUFNAGEL, S.B., KLEIN, T.E., KORF, B.R., MCKELVEY, K.D., ORMOND, K.E., RICHARDS, C.S., VLANGOS, C.N., WATSON, M., MARTIN, C.L., MILLER, D.T. "Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics". *Genetics in Medicine* 2017;19(2):249-255.

12. LEUNG, L. "Incidental Findings in Neuroimaging: Ethical and Medicolegal Considerations". *Journal of Neuroscience* 2013;2013:439145. doi: 10.1155/2013/439145.

13. MAILMAN, M.D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., IYABARI, L., LEE, M., SHAO, Y., WANG, Z.Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J., SHERRY, S.T. "The NCBI dbGaP database of genotypes and phenotypes". *Nature Genetics* 2007;39(10):1181-6.

14. NATIONAL INSTITUTES OF HEALTH. (2012). *1000 genomes project data available on Amazon cloud*. http://www.nih.gov/news/health/mar2012/nhgri-29.htm. Accessed May 30, 2017.

15. RAISARO, J.L., TRAMÈR, F., JI, Z., BU, D., ZHAO, Y., CAREY, K., LLOYD, D., SOFIA, H., BAKER, D., FLICEK, P., SHRINGARPURE, S., BUSTAMANTE, C., WANG, S., JIANG, X., OHNO-MACHADO, L., TANG, H., WANG, X., HUBAUX, J.P. "Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks". *Journal of the American Medical Informatics Association*. 2017 Feb 20. doi: 10.1093/jamia/ocw167.

16. SHRINGARPURE, S.S., BUSTAMANTE, C.D. "Privacy Risks from Genomic Data-Sharing Beacons". *American Journal of Human Genetics* 2015 Nov 5;97(5):631-46. doi: 10.1016/j.ajhg.2015.09.010.

17. STEPHENS, Z.D., LEE, S.Y., FAGHRI, F., CAMPBELL, R.H., ZHAI, C., EFRON, M.J., IYER, R., SCHATZ, M.C., SINHA, S., ROBINSON, G.E. "Big Data: Astronomical or Genomical?" *PLoS Biology* 2015;13(7):e1002195.

18. WADMAN, M. "James Watson's genome sequenced at high speed". *Nature* 452 (7189): 788–788. (2008) doi:10.1038/452788b.

19. WETTERSTRAND, K.A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: www.genome.gov/sequencingcostsdata. Accessed 29/03/2017.

20. WRIGHT, C.F., MIDDLETON, A., BURTON, H., CUNNINGHAM, F., HUMPHRIES, S.E., HURST, J., BIRNEY, E., FIRTH, H.V. "Policy challenges of clinical genome sequencing". *BMJ* 2013;347:f6845.

21. WRIGHT, C.F., MIDDLETON, A., PARKER, M. "Ethical, legal, and social issues in clinical genomics". In: Kumar D and Eng C (eds.) *Genomic medicine: principles and practice*. Oxford University Press 2015, New York, USA.

Fecha de recepción: 2 de junio de 2017
Fecha de aceptación: 24 de junio de 2017