Journal of
Clinical Research & Bioethics

Meller, J Clin Res Bioeth 2015, 6:6
DOI: 10.4172/2155-9627.1000249

Review article Open access

Addressing Benefits, Risks and Consent in Next Generation Sequencing Studies

Meller R[*]

Translational Stroke Program, Neuroscience Institute, Morehouse School of Medicine, Atlanta, USA

[*]**Corresponding author:** Meller R, Neuroscience Institute, Morehouse School of Medicine, 720 Westview Dr SW, Atlanta, GA, 30310, USA, Tel: 4047565789; Email: rmeller@msm.edu

Received date: October 13, 2015; **Accepted date:** December 11, 2015; **Published date:** December 14, 2015

Abstract

The sequencing of the human genome and technological advances in DNA sequencing have led to a revolution with respect to DNA sequencing and its potential to diagnose genetic disorders. However, requests for open access to genomic data must be balanced against the guiding principles of the Common Rule for human subject research. Unfortunately, the risks to patients involved in genomic studies are still evolving and as such may not be clear to learned and well-intentioned scientists. Central to this issue are the strategies that enable human participants in such studies to remain anonymous, or de-identified. The wealth of genomic data on the Internet in genomic data repositories and other databases has enabled de-identified data to be broken and research subjects to be identified. The security of de-identification neglects the fact that DNA itself is an identifying element. Therefore, it is questionable whether data security standards can ever truly protect the identity of a patient, under the current conditions or in the future. As Big Data methodologies advance, additional sources of data may enable the re-identification of patients enrolled in next-generation sequencing (NGS) studies. As such, it is time to re-evaluate the risks of sharing genomic data and establish new guidelines for good practices. In this commentary, I address the challenges facing federally funded investigators who need to strike a balance between compliance with federal (US) rules for human subjects and the recent requirement for open access/sharing of data from National Institute for Health (NIH)-funded studies involving human subjects.

Keywords: Generation sequencing; Consent; Human subject research

Introduction

The release of the human genome sequence has given rise to new approaches for medicine. The linking of genomic information and patient health information, coupled with advances in computing and informatics techniques, has the potential to herald a new era of molecular medicine. Advanced medical diagnostics, prognostics and therapeutic tailoring are all potential benefits from such discoveries. However, the massive amount of data collected and the potential for identification pose risks to patients who agree to participate in genomic studies. Central to this following commentary is the fact that attempts to de-identify patient data may be futile given that a patient's genomic information (DNA) may be the "ultimate identifier". Here I consider the potential benefits and risks that must be balanced in genomic studies.

Benefits of Genomic Studies for Medicine

Technological advances have facilitated the development of cheap, high-throughput sequencing systems, now capable of sequencing a human genome for $1000 [1,2]. Advances in bioinformatic techniques and powerful computational methods to collate and mine biologically important information from genomic datasets have kept pace with this data explosion. As the use of bioinformatic techniques becomes more common in clinical research, and genetic and genomic information is linked to patient health records, the power of these techniques increases. This is a classic view of Big Data, whereby genomic data is linked to health information to generate more advanced and subtle health information for an individual patient. Such initiatives are currently being launched on both national and international scales [3]. As such, the benefits to society can easily be determined based on the health benefits from NGS technologies.

The potential for societal benefit from genomic studies is very high, although still being realized [4]. Advanced molecular diagnostics may revolutionize the identification (diagnosis) of disease. Examples where such advances have increased our understanding of disease include understanding the role of mutations in cancer pathologies resulting in a molecular basis for cancer diagnosis [5], understanding the role of genomic abnormalities in neurological disorders [6], the development of pre-natal screening methods using maternal blood samples [7], and enhanced diagnostic tools for acute neurological injury [8]. Potentially, patients who are at risk of developing diseases can be identified, and prevention programs specifically tailored to a patient's genetic profile can be devised. It may also be possible to establish a measure of disease and healthiness.

The use of genomics to predict patient response to therapy may enable the individualization of medicine. Genetics or genome profiles could be used to identify therapeutics that offer a patient the greatest benefit and least risk. This is no longer theoretical; KRAS signalling mutations have been used to guide therapy [9]. This customized or precision medicine-based therapeutic strategy may radically change current approaches for rationalized therapy for many disorders, not just cancer.

In order to build a detailed knowledge of how genomics profiles affect susceptibility to disease and therapeutic responses, we need to analyze large genomic data sets. Such data sets can only be created or collated by large multi-center projects or by sharing data. The goal behind data sharing is to maximize the knowledge gained from a given study. Data from a given project may by useful in other related projects. For example, studies of the mechanisms of cell death in cancer have guided studies of cell death in neurological disease, where a reduction in cell death is the goal. In addition, by sharing data, researchers can combine data, potentially increasing the power of a given study. Data sharing is important when multi-center initiatives generate data; common data handling and data standards enable the harmonization of resources (for example, standards proposed by the Global Alliance for Genomics and Health). From a financial standpoint, data sharing also reduces unnecessary replication, while allowing maximal access to given datasets. Therefore, as multi-center initiatives in genomics develop, the shared data resources available to researchers will become invaluable. However, data sharing must be balanced against the protection of privacy [10].

NIH Genomic Data Sharing Policy

To maximize the benefits of large (complex and expensive) datasets, the NIH recently requested that all genomic data generated from NIH-funded projects be made available to the community under shared or controlled access. Given the potential benefits to society (see above) and the wealth of information available to researchers, one can clearly see how shared access to such data will assist in the biomedical research of disease.

NIH (Not OD-13-119) set forth the planned NIH genomic data sharing policy such that all NGS data generated by federal funds be shared [11]. A few points in the policy are cause for concern. First, while removal of personal identification information from the data set is requested, the guideline is not as detailed as the HIPPA standard for de-identification. Second, the policy states that "prior to data submission, traditional identifiers such as name, date of birth, street address, and social security number should be removed. The de-identified data are coded using a random, unique code to protect participant privacy [11]." However, there is a fundamental assumption that the sequence data is not itself an identifier or a potential source of data for identification. Recent studies challenge this assumption.

Identifying personal genomes by surname inference

Patients can be re-identified using DNA sequence information. In a study by Grymek et al. reported in Science in January 2013, the surnames of de-identified human subjects were recovered from chromosome Y tandem repeat data and genealogical databases [12]. The authors used freely available public Internet data to perform the analysis. The Grymek study noted that consent forms identified data breech as a security risk, that re-identification was not prevented, that privacy could not be guaranteed, and that future techniques might be able to identify the subjects [12]. The study concluded that surname inference from public databases is feasible, putting the privacy of current de-identified data sets at risk. One of the more troubling aspects of the Grymek study is that one of the co-authors, who developed the software used in the study, decided to perform the study because "they could not resist trying" [13]. This suggests a potentially worrying lack of oversight or regulation.

However, the idea that DNA can identify a participant is not new. In 2006, public records and death records enabled a researcher to identify people based on familial pedigree studies of genetic markers [14]. Use of genealogical databases has resulted in the parental identification of adopted children [15]. In addition in 2008 and 2011, information from GWAS studies enabled participant identification [16,17]; the more information about mutations in a person's genome the greater the risk of identification. While the identification of a family or parent of a child may have damaging consequences, far more information can be extracted from genomic information.

The significance of the Grymek study, with respect to its relevance to the US population, has been challenged. Indeed, some groups believe that the study was published to appeal to the media rather than to inform the public of real re-identification risks [18]. The surname inference algorithm was tested first against a well-known individual who had published his genome (Craig Venter) and then against a well-defined population (Utah residents of European ancestry associated with the Mormon Church) [12]. Of note, the data set had previously been shown to be vulnerable to re-identification. Critics also opine that the re-identification study was only possible in an academic department with appropriate resources [18]. These opinions fail to acknowledge the fact that as more information becomes available via databases and online repositories, further inferences will be possible, especially when other pieces of information can be ascertained from public sources and built into the model.

One consideration that may not be immediately obvious to a genomic study participant is the power of Big Data approaches to obtain information from non-traditional sources, such as social media, for example [19]. Big Data approaches are being evangelized and debated for their potential benefits to society. However, less consideration has been given to the privacy concerns associated with Big Data. Facebook is estimated to have over one billion users. Such sites, thus provides a wealth of personal information that can be mined and used for identification purposes along with a few demographical identifiers. Newer biometric devices also give access to health information. The number of internet connected devices is predicted to top 50 billion in 2020 [20]. Therefore, social media and online resources contain a wealth of data that could be used to assist in patient identification.

While Big Data may not immediately pose a threat to re-identification, there is cause for concern in the future. Techniques and algorithms to mine large data sets have been developed, and similar algorithms might be adapted to mine publically available data for similar inferences and personal data. An issue with Big Data is the vagueness of the term and processes involved, as well as the lack of transparency around what Big Data is and those performing the research. Here we define Big Data as a process of sifting and assimilating disparate data to obtain meaning, but other definitions are also used. Recent revelations of the scope of Big Data collection and infrastructure infiltration by the National Security Administration (NSA) in the USA has resulted in fears for privacy that are hard to overcome. A consideration of the risk to benefit of Big Data is beyond the scope of this review, but one must be mindful of the power of such large computational systems to infer from disparate and varied social media sites as well as from more identifiable sources of health-associated information.

Recent studies suggest that younger generations are more open than older generations when it comes to sharing personal information [21]. As direct-to-consumer DNA sequencing services (such as 23andMe and genealogy services) become available, the open nature of the former demographic may result in open access online sharing of personal information [22]. The consequences of open sharing are probably not acknowledged, but the risks should be clearly identified. This is especially pertinent for participants in genomic studies, as the ability to identify a patient using information about gene mutations becomes possible and Big Data methodologies become more refined.

The challenge to the community is to determine what changes to the rules and regulations regarding what one can do with genomic information are needed. In the future, it is likely that participants will be identifiable from their genetic/genomic information. As such, it is perhaps unrealistic to attempt to guarantee the privacy or anonymity of participants. It is recommended that the public be better educated during the consent process about the risks of identification.

What is the damage of releasing personal genomic data and data from sequencing?

When considering the risks to a participant in a genomic study, one should consider why re-identification would be an issue for a participant. Clearly, the identification of a participant and their genomic information, with information about mutations associated with disease, would enable a number of inferences about the person, their current health and their future health. Notwithstanding the academic question of whether one can re-identify a given patient in a study, other groups may have an interest in such data (this is not an exhaustive list):

Insurance companies: Federal rules and laws prohibit medical insurance discrimination based on genetic data. However, the enforcement of such polices may be problematic, and violation of such rules may be hard to prove. While health insurance cannot be denied or rates increased because of genetic disorders [23,24], other forms of insurance, such as long term care policies, are not subject to these laws. This may need to be addressed in the near future.

Employment: Some opinion posts suggest that employers may discriminate on the basis of finding certain diseases or personality traits. This is prohibited under the 2008 Genetic Information Non-discrimination Act (GINA). For example, certain APO4E variants are associated with a markedly higher risk of Alzheimer's disease. Would an employer hire someone with a heightened chance of Alzheimer's or other costly disorder? In 2010, a patient filed a GINA lawsuit against her employer for dismissal subsequent to a BRACA1-positive test result [25]. This case was settled out of court with a non-disclosure agreement.

Law enforcement: A commonly cited, albeit theoretical concern, is the ability to match genomic data to DNA fingerprinting data. DNA fingerprinting involves the identification of DNA microsatellite patterns. Such sequences could be interpreted from an annotated genome and then used to search DNA fingerprinting files. This remains a theoretical concern at this time.

Unknown family members: Genealogical data has already been useful in searches for parents, either by adopted children or children conceived through artificial donor insemination. A recent case in Kansas, in which a donor was sued for child support, highlights some of the potential issues associated with this situation [26]. The case was prosecuted by the state, even though the donor had signed away his paternal rights. In addition to the obvious cost, confrontations between sperm donors and potential children could cause substantial personal/social issues.

Criminals (blackmail): A genetic trait or vulnerability could be used to extort money. In addition, one could easily envision a scenario in which paternity and evidence of communicable disease could be prosecuted as illegal activity. No cases of genetic data being used to blackmail a person have yet surfaced.

Other: Another concern with the release of identifiable genetic/genomic information is the potential consequence for other family members. The recent controversy surrounding the sequencing and publication of the HeLa cell genome [27] and the effect on the family of Henrietta Lacks resulted in a new NIH policy on genomic studies of HeLa cells (NOT-OD-14-08). Few diseases have such direct a correlation between risk and disease, when the disease is not already apparent or predictable. For example, Huntington's chorea has a strong identifiable familial linkage, and BRCA mutations are associated with a high risk of certain cancers, but other progressive diseases are not as easily identifiable in genetic data at this time. The risk in this situation is that a person may not possess the knowledge or tools to correctly interpret the data or the risk, leading to unfortunate consequences. James Watson discovered that he had a BRCA mutation. Fortunately, he obtained genetic counselling before announcing the discovery to family members. The mutation was not associated with disease, therefore obviating the need additional testing. However, this is a clear example of the potential for harm from limited data. Indeed, one recent post suggests that only negative information is obtained from DNA mutations [28]. Clearly, there is concern is for family members of identified people with genetic indicators that predict a higher likelihood of diseases that have expensive and/or prolonged therapies.

Disorders with a genetic basis can be predicted directly from genetic/genomic data (types 1-3). Once a potential candidate for a disease is identified, the challenge is to identify the person. It is hoped that legislation will thwart such use of genetic data, either by limiting the use of genetic information for insurance calculations or prohibiting re-identification [18].

What other information can be obtained from genomic data?

Most of the risk associated with genetic/genomic studies involves the identification of human diseases with a genomic basis and subsequent patient identification. A second and perhaps less obvious risk is the potential to identify infectious diseases from genomic/genetic data sets. The current data policy enables the identification of non-human DNA from other micro-organisms such as bacteria and viruses. The NIH policy states that both aligned and unaligned reads are to be submitted (Level 2 data file (e.g. binary alignment matrix (BAM) or equivalent). A BAM file is an indexed data file that contains the sequence and genomic location of a sequenced section of DNA. The request for unaligned reads is a potential for concern, because these data contain information on DNA that does not align to the human genome, for example viral or bacterial DNA information. There may be risks if such information can be matched to a given patient. Such a concern is not theoretical. It was recently shown that microbes associated with sexually transmitted infections (STIs) could be identified in asymptomatic patients [29]. In this study, S16 ribosomal RNA patterns were used to identify various STIs once the human sequences were removed from the data set. This in essence is the same data as "unmapped DNA", which is required in NIH data submissions [11]. Therefore, similar approaches could be used to identify bacterial or viral genetic information from a patient's tissue. For example, evidence of hepatitis C or human papilloma virus might be used to increase the cost of health insurance due to the long-term potential cost of hepatitis and HPV-induced cancers. A second and perhaps more nefarious use of such data would be to blackmail or extort a patient to prevent public release of this information. In addition, a social stigma is frequently associated with STIs, which could affect patient decisions in consenting to infectious disease research [30].

The identification of infectious diseases requires raw sequencing data (level 1 data) or unaligned level 2 data. NIH policy suggests depositing the unaligned reads in the database. This would make searching easier, because the data will have already been filtered for human genomic sequences, thereby reducing the computational power required for such studies.

However, while we see the risks in such information, it is clear that the information is also beneficial to research. A number of projects have focused on the characterization of the microbiome, a complement of microbial organisms in different body compartments that influence human health. This area of research is currently a focus for the NIH, because it has the potential to change how we view disease. Furthermore, the identification of viral infections may assist epidemiological infectious disease research. Hence, these data have utility, but a patient should be able to assess the risk of consenting to such a study.

Consideration of other "omic" data sets

Considerations of the risks associated with genomic information mostly focus on DNA sequencing. However, the sequencing of all gene products should be subject to a similar risk assessment. The methodology used by Grymek to identify patients from tandem repeat analysis [12] may not be applicable to RNA-seq, because these elements may not be transcribed into RNA. However, because RNA is directly transcribed from DNA, it is easy to infer the DNA sequence from the RNA sequence. Furthermore, it may be easier to identify a person from RNA than from DNA. Protein coding regions of DNA/RNA show low rates of single nucleotide polymorphisms (SNPs), whereas non-coding regions of RNA, such as intergenic and intronic RNA, have higher rates of base differences/SNPs [31]. Therefore, unique differences in these regions could be identifiers. Protein-based sequencing (proteomics) may also be subject to similar considerations; both proteomics and RNA sequencing can identify bacterial or viral products. Therefore, it is prudent to consider genomic product sequence information with the same rigor as DNA sequence information.

How can we reduce the risk to patients and maintain potential benefits?

What is a potential solution? There are risks for a participant in a genomic study, but there are potential benefits too. Clearly, a consent form will have to address both. How to draft a consent form so as to describe the above risks while balancing them against potential benefits is an area of debate [10,32]. Two points seem central to the discussion. The first and central point is that DNA is an identifier; the removal of names and other protected health information/identifiers may be rendered mute by this fact. Therefore, claims of privacy and confidentiality must be grounded in a forward-looking perspective. Second, many debates have focused on informatics data release posing a minimal risk of harm. I contend that this may not be the case, and that provisions within the framework for determining the minimal risk of genomic studies may not be appropriate.

Previous approaches have used a broad consent process, whereby the participant is generally informed of the risks of participating in research, with specific risks associated with a particular type of research not included. It has been argued that a broad consent can be used if the details of governance are laid out to the participant [33]. This argument has been rejected by others [34], and more dynamic/continuous consent processes have been proposed [35,36]. This approach has many merits, including the ideal that research participants be kept informed and involved in the research process. A prospectively obtained broad consent may be limited in its ability to accurately spell out the potential future risks of identification or future risks in information revealed in a genomic dataset. Since the autonomy of a subject participating in a research program is clearly challenged by a broad consent form, a broad consent no longer appears the appropriate mechanism for genomic studies. However, counter to this argument, a highly detailed consent form is also unlikely to result in a "reasonably informed person".

The recent proposed changes for NIH-funded genomics studies recommend that future IRB consent forms include language informing patients that their genomic data will be shared for future research. Sharing in open access repositories requires explicit consent [11]. An allowance is made to enable controlled data access, especially if informed consent was not obtained. Such limited access must be outlined in the grant submission. "Upon request, limited controlled access for submitted data can be requested. In such a situation the institution can restrict what the data will be used for. The requestor must then have their project approved by NIH and are subject to restrictions on what they can do with the data." Notably, in surveys of attitudes to data sharing patients express a greater level of comfort when the data is restricted to approved researchers [30].

In devising the consent form for genomic studies, a decision regarding the potential use of the data might be necessary. This is clearly easier for a defined genomic study than for a biobank/repository consent form. A recent presidential review committee strongly suggested that Whole Genome Sequencing studies should not be performed without explicit consent (a position this author is in agreement with) [37]. As such, opt out should not be used when there is the potential for future sequencing. Furthermore, consent forms need to include language about the risks of genome sequencing and determine what data sets are released. NIH data sharing policies require sharing of level 2 or human genome aligned BAM files (Table 1). The question then remains: do we want to publically release the data in an open access format, or should all human data be on a restricted access policy? The public seem more comfortable with restricted access release [30]. A tiered consent form could address these options, or dynamic consent could inform a participant of the potential uses of the samples. To increase a participant's trust in a project, details in a consent form may also address how restricted access will be governed [33].

NIH polices request the release of unaligned data. The release of unaligned data may have great utility, for example identifying infection in a population or microbiome changes associated with a clinical condition or disease. Given the risks identified above, a separate explicit consent may be considered if data will also be aligned to non-human reference genomes. Perhaps re-analysis of data to reference genomes not explicitly referenced in the original consent procedure protocol should require an amendment of the original IRB protocol and even re-consent of the participant. A second issue not discussed above is the requirement for reporting if evidence is gained for certain transmittable diseases. As NGS technologies become used for clinical diagnostics, this may soon require further consideration (Table 1).

Data Levels	File type	Description	Information contained in data
Level 0	Image file (.Tiff)	Raw data obtained from the DNA sequencing system. Depends on the methodology of the sequencer, but typically either a series of images, or a reading of individual bases	Limited raw data. This usually requires additional information to decode the individual base sequences and assign them to a given sample.
Level 1	Raw read files (fastq, xsq, ssf)	Cleaned raw read files, where raw data is converted into a series of DNA bases and the quality of the DNA sequence is determined.	These files contain all of the raw DNA sequence information of a patient. However, they are not yet aligned with the host (human?) genome. Also contains non-human DNA sequences from non-human, i.e. viral, DNA present in the sample.
Level 2	Aligned data files (BAM, SAM, CRAM)	DNA reads are aligned to their corresponding location on the host (human Genome). DNA reads that do not align with the host can be re-analyzed for alignment to other genomes (i.e. bacterial or viral). For RNA, this includes RNA expression profiling.	The alignment of files enables rapid identification of the species the DNA comes from and its location within the genome. This enables rapid identification of mutations, etc.
Level 3	Analysis software data output (proprietary format, or spreadsheet/ database file)	Analysis of DNA sequences to identify base mutations expression patterns or other features of the data set.	These files contain summaries of the "features" of the data, for example mutations, or gene expression values. Typically generated from the analysis software.
Level 4	Sequence data aligned to phenotype (proprietary format, or spreadsheet/ database file).	Summarized data, with key features identified along with clinical features (phenotype)	These files combine the "features" with the clinical data (phenotype).

Table 1: Data levels obtained in DNA sequencing studies. DNA sequencers function by identifying the sequence (base composition) of short reads of DNA. These reads are then aligned with a reference genome. Therefore, data obtained from sequencers typically comprises raw image files (Level 0), raw read files (Level 1) aligned and cleaned data files (Level 2), analyzed data files (Level 3), and data correlated to phenotype (Level 4).

One central theme here is that DNA is an identifier. However, additional pieces of information can assist in the identification of a patient. De-identification strategies have focused on the removal of four critical identifiers: name, date of birth, address and social security number. The Health Insurance Portability and Accountability of 1996 Act (HIPAA) suggests scrubbing of 19 potential determining factors. In the Gymek study [12], dates of birth were inferred from ages. As such, the question remains: if ages were stratified in 5-10 year blocks, rather than in 1-year divisions, would the data still have as much utility to a researcher? If data stratification is performed using age, 10-year blocks may be sufficient to at least make attempts of re-identification based on age harder, but not impossible with enough information. Therefore, it would seem appropriate to release stratified rather than absolute ages of participants.

Studies that have been able to re-identify patients have relied on a reference [12,16,17]. Guidelines to assist participants in remaining un-identified may be considered and suggested. Indeed, many studies of re- identification suggest we must educate patients to the risk of such studies. The following are potential inclusions to consent forms:

Explain whether the genomic data will or will not be released to open access or restricted access database repositories

Tier the consent form to include human and non-human (infectious disease) genomic data.

Explain that absolute privacy cannot be guaranteed in the future

Explain that the patients may compromise their privacy via the use of genealogical studies etc. and other social media platforms

Final Comments

The Big Data revolution is upon us, and a key contributor to this movement is the largescale acquisition of genetic information linked to clinical/health information [3]. As we move into a new era of data science and information-driven healthcare, we must determine best practices to ensure minimal risk to patients and maximal societal benefit from the data generated. Public trust in science is paramount for the success of such studies. While guidance for scientists involved in such research is becoming clearer, the area is quite complex, combining elements of ethics and risk assessment, as well as research science [38]. We will not be able to foresee all risks regarding information release, but with some foresight, we can attempt to reduce the risks.

Acknowledgment

I wish to acknowledge the assistance of Profs PR MacLeish, PhD and RP Simon, MD and A Pearson for helpful discussions in developing this article, and Dr M Johnson for editing. R Meller is supported by NIMHD U54 MD007588, (NINDS) R01NS59588. The views represented in this article are those of the author and do not represent the views of the Morehouse School of Medicine, NIH, NINDS or NIMHD.

References

1. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155: 27-38.
2. Hayden EC (2014) Technology: The $1,000 genome. Nature 507: 294-295.
3. Eisenstein M (2015) Big data: The power of petabytes. Nature 527: S2-4.

4. Hudson K (2008) The health benefits of genomics: out with the old, in with the new. Health Aff (Millwood) 27: 1612-1615.
5. Burghel GJ, Hurst CD, Watson CM, Chambers PA, Dickinson H, et al. (2015) Towards a Next-Generation Sequencing Diagnostic Service for Tumour Genotyping: A Comparison of Panels and Platforms. Biomed Res Int 2015: 478017.
6. Martin HC, Kim GE, Pagnamenta AT, Murakami Y, Carvill GL (2014) Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. Hum Mol Genet 23: 3200-3211.
7. Landau YE1, Lichter-Konecki U2, Levy HL3 (2014) Genomics in newborn screening. J Pediatr 164: 14-19.
8. Dykstra-Aiello C, Jickling GC, Ander BP, Zhan X, Liu D (2015) Intracerebral Hemorrhage and Ischemic Stroke of Different Etiologies Have Distinct Alternatively Spliced mRNA Profiles in the Blood: a Pilot RNA-seq Study. Transl Stroke Res 6: 284-289.
9. Prenen H, Tejpar S, Van Cutsem E (2010) New strategies for treatment of KRAS mutant metastatic colorectal cancer. Clin Cancer Res 16: 2921-2926.
10. Kaye J (2012) The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet 13: 415-431.
11. Tabak L, Final NIH Genomic Data Sharing Policey. Federal Register, 2014. 79(167): p. 51345-51354.
12. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. Science 339: 321-324.
13. Kolata G (2013) Web Hunt for DNA sequences leaves Privacy Compromised, The New York Times, New York.
14. Malin B (2006) Re-identification of familial database records. AMIA Annu Symp Proc.
15. Lunshof JE, Chadwick R, Vorhaus DB, Church GM (2008) From genetic privacy to open consent. Nat Rev Genet 9: 406-411.
16. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4: p.e1000167.
17. Masca N, Burton PR, Sheehan NA (2011) Participant identification in genetic association studies: improved methods and practical implications. Int J Epidemiol 40: 1629-1642.
18. Michelle M (2013) Public Policy Considerations for Recent Re-Identification Demonstration Attacks on Genomic Data Sets: Part 1 (Re-Identification Symposium).
19. Musaev A, Wang DE, Pu C (2015) LITMUS: A Multi-service Composition System for Landslide Detection. IEEE Transactions on Services Computing 8: 715-726.
20. Danova T (2013) Morgan Stanley: 75 Billion Devices Will Be Connected To The Internet Of Things By 2020. Business Insider.
21. Anderson JQ and Rainie L (2010) Millennials will make online sharing in networks a lifelong habit. Pew Research Center: Washington.
22. Knoppers BM (2010) Consent to 'personal' genomics and privacy. Direct-to-consumer genetic tests and population genome research challenge traditional notions of privacy and consent. EMBO Rep 11: 416-419.
23. Centers for Medicare & Medicaid Services (CMS), Department of Health and Human Services (HHS) (2014) Patient Protection and Affordable Care Act; exchange and insurance market standards for 2015 and beyond. Final rule. Fed Regist 79: 30239-30353.
24. Sarata AK, Jones NL, Staman J (2011) The Genetic Information Nondiscrimination Act of 2008 and the Patient Protection and Affordable Care Act of 2010: Overview and Legal Analysis of Potential Interactions. Congressional Research Service.
25. Lehmann-Haupt R (2010) GINA Violated? Employee Claims Genetic Test Result Got Her Fired.
26. Press A (2013) Kansas: Sperm Donor Is Ordered to Pay Support.
27. Landry JJ, Pyl PT, Rausch T, Zichner T, Tekkedil MM, et al. (2013) The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda) 3: 1213-1224.
28. Lewis R (2014) James Watson On “Genetic Losers”.
29. Nelson DE, Van Der Pol B, Dong Q, Revanna KV, Fan B, et al. (2010) Characteristic male urine microbiomes associate with asymptomatic sexually transmitted infection. PLoS One 5: p.e14116.
30. Robinson JO, Slashinski MJ, Chiao E, McGuire AL (2015) It depends whose data are being shared: considerations for genomic data sharing policies. Journal of Law and the Biosciences Press (Advance Access).
31. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 39: 7058-7076.
32. McGuire AL, Achenbaum LS, Whitney SN, Slashinski MJ, Versalovic J, et al. (2012) Perspectives on human microbiome research ethics. J Empir Res Hum Res Ethics 7: 1-14.
33. Sheehan M (2011) Can Broad Consent be Informed Consent? Public Health Ethics 4: 226-235.
34. McGuire AL, Beskow LM (2010) Informed consent in genomics and genetic research. Annu Rev Genomics Hum Genet 11: 361-381.
35. Kaye J, Whitley EA2, Lund D3, Morrison M, Teare H, et al. (2015) Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 23: 141-146.
36. Steinsbekk KS, Kåre Myskja B, Solberg B (2013) Broad consent versus dynamic consent in biobank research: is passive participation an ethical problem? Eur J Hum Genet 21: 897-902.
37. Privacy and Progress in Whole Genome Sequencing (2012) Presidential Commission for the Study of Bioethical Issues, Washington DC.
38. Ebbesen M, Sundby A, Pedersen FS and Andersen S (2015) A Philosophical Analysis of Informed Consent for Whole Genome Sequencing in Biobank Research by use of Beauchamp and Childress’ Four Principles of Biomedical Ethics. J Clin Res Bioeth 6: 244.