Security of human genomic data

Fangfang Qu

Abstract

Precision medicine, tailoring of medical treatment to the individual characteristics, is the future of healthcare. Genomic based tests as the trigger of precision medicine revolution and the associated genomic data in clinical practice are blowing up in recent years as the price goes down and more biomedical supports on the data interpretation. With the specialty and uniqueness of this kind of data, security and protection on these data can hardly be overemphasized. This paper will first discuss the specialty of genomic data and potential harmful consequences of a big data breach. Second, the reasons for vulnerabilities of genomic data will be analyzed. Furthermore, the paper will examine possible strategies for individual users, biobank, researchers and data centers to improve genomic data protection.

1. Introduction

Precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." [1] In contrast to a one-size-fits-all approach, Precision medicine tailors treatment and prevention strategies according to the individual characteristics. The rapid advancement in next generation sequence of genomic information has accelerate the revolution of precision medicine. With a total of about 3 billion DNA base pairs, the human genome embeds essential information about the development, physiology, pathology and evolution of human being. Uniqueness among individuals' genetic composition is presented in genetic markers such as short tandem repeats loci (STR) and genetic variants including difference in a small size of bases (SNP) and a larger segment of DNA region (structural variant) [2].

Next generation sequencing (NGS), determining the exact order of the base pairs of DNA segment, is extremely flexible for sequencing a whole exome (all of the coding regions of DNA), a whole genome (entire DNA sequence including both coding and noncoding regions). For a typical workflow of NGS [3], generally, the DNA sample collected from hair, skin, blood saliva or tissue is extracted, fragmented. The fragment library are prepared by end-repairing, adaptors ligation and amplification, then subject to sequencing in a massively parallel mode. Large amount of short reads were generated by the sequencing platform. Up to this point, the chemical DNA is converted to digital DNA. Post sequencing analysis involves in data storage, transmission, standard bioinformatics process with different bioinformatics tools, and data sharing among organizations, researchers, clinician, individual user. The vulnerability of these data could be results from each of above processes with regard to cyber security.

Genomic data not only enables uniquely identification of individuals, it also provides information about genetic predisposition of an individual for common diseases, familial traits, carrier status for inherited diseases, and adverse reactions to common drugs. As respect to its special features, genomic data requires careful handling, new regulations to ensure data security and more education about data privacy in individual level.

2. To the Community:

A genome represents a treasure of highly personal and sensitive information. As the cost of full genome sequencing goes down and the interpretation of genomic data becomes more meaningful and useful, large numbers of people start to generalize their digitized genomes. Prioritizing security of personal genomic data becomes extremely important. In this paper, I discuss the speciality of genomic data and the privacy risk of genomic data leakage, and analyze possible causes of insecurity and vulnerabilities of genomic data. The actions of all related parties for better protecting genomic data are further discussed.

3. specialty of genomic data

Genomic data is high sensitive and has its own speciality comparing to other sensitive data. At first, the profile of a germline genome is almost immutable over a person's life time, with a few exceptions that are caused by sequencing inaccuracy or aging induced shortening of the ends of DNA strands [4]. While changing the credit card number can solve the issue caused by losing a credit card, there is no possible way to update your genome information. The security risk of this immutable future it that an individual may be fraud by a life-long time because of genomic data leakage.

Secondly, an individual's genomic data represent not only his/her own genetic information, but also their ancestors, descendants, and siblings because of high similarity in their genome [5]. Thus, even people who never have their genome sequenced could be compromised by blood relatives with their genome sequenced. Meanwhile, genomic data are highly valuable because of the broad application of one's digitized genome in health and behavior such as paternity testing, inference of adverse reactions to common drugs,ancestry tracing and disease screening.

Furthermore, along with the fast growing of scientific knowledge by intensive studies involved in genome research, more new and accurate interpretation of one's genomic data may be revealed in the future. You may never know what will be possible to do with your genome a decade or two from now. Consequently, with all the special features, genomic data security and privacy is a very timely and important subject.

4. Privacy Risk of genomic data breach

Anonymization and aggregate genomic information is done by stripping individual's personal details including name and contact information and aggregated with the genomic information of others. However, it has shown that standard anonymization techniques are ineffective with genomic data and, as a consequence, anonymized genomes can be used to recover the identities of the individuals which is called re-identification [6]. Re-identification is one of the most serious privacy risk of human genomic data. In 2014, Yaniv Erlich et. al [7] summarized genetic privacy breaching strategies, and illustrate the routes they used to re-identify a person using the anonymous genomic data with with some basic demographic information. As shown in figure 1, at first the sex was inferred by the sex chromosomes, then the metadata was used to find the state and the age of this person, after that the person's surname was recoveried by a public available genetic-genealogy database, Ysearch. Furthermore, a public record search engines such as PeopleFinders.com was used to generate a list of

potential individuals. Finally, social engineering or pedigree structure can be used to triangulate the person. This pioneer study highlights the importance of protect the genetic privacy of the data originators.

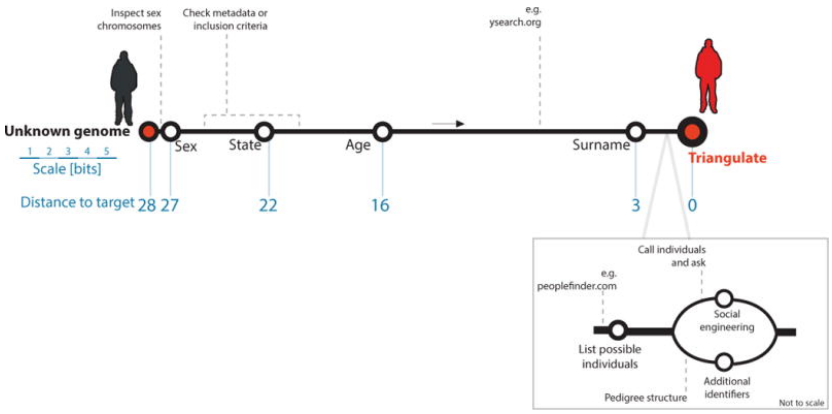

Figure 1. A possible route for identity tracing using anonymous genomic data [7].

The leakage of genomic information may cause serious results by genetic discrimination or blackmail. The regulation law or policy for protecting the privacy of individuals' genomic data are far from perfect. Since the Health Insurance Portability and Accountability Act (HIPAA) does not cover anonymized data, in 2007, the U.S. adopted the Genetic Information Nondiscrimination Act (GINA), which prohibits certain types of discrimination in access to health insurance and employment. However, GINA does not apply to life insurance or the military; thus a long-term-care insurance company may penalize someone for a genetic predisposition to certain disease such as cancer or Alzheimer's disease [8]. A true story about genetic discrimination was shared by Dr. Lindor [9]. One of her cancer patient's grandchildrens was rejected by the U.S. army for an application of helicopter pilot, because she carries the same mutation as her grandparent for the specific type of cancer.

5. Cause of insecurity and Vulnerabilities

Clear text Transmission of Sensitive Genomic Information

Since data is often shared among researchers and sequencing facilities, One common vulnerability is that genomic data is transferred using FTP through internet or internal organization networks. By searching on shodan.io [10] using "FTP genome", "FTP bioinformatics", or "FTP sequencing", I have got a list of universities or institutes running FTP server to transfer human genome sequencing data, as shown in figure 2. Without use a secure channel, such as SSL, to exchange sensitive information, it is possible for an attacker with access to the network traffic to sniff packets from the connection and uncover the sensitive data. And this type of attack is not technically difficult.

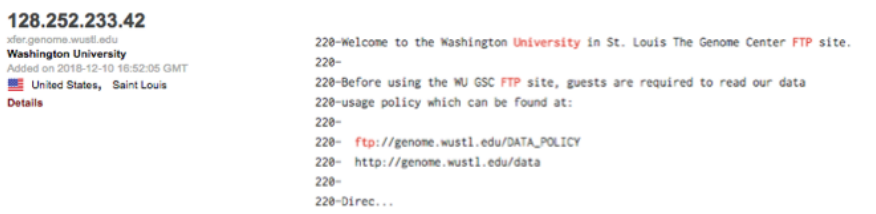

Figure 2. FTP site of genome center exposed by shodan

Vulnerability in bioinformatic tools

As rapid improvement of next generation sequencing technology, it has speed up the development of bioinformatics tools and pipelines applied for sequencing analysis. Typically, these Bioinformatics pipelines for next generation sequencing are often developed by small research groups and open source. However these tools are not always written with security in mind. One group evaluate the software security in a wide set of DNA processing programs and they found that existing biological analysis programs written with C have a much higher frequency of insecure C runtime library function calls such as strcpy, sprintf or strcat [11]. These function calls may cause buffer overflow and program crashes and subsequently converted to exploits.

Vulnerability in Shared Databases

Commonly NGS results are analyzed and shared in a supply chain which includes bioBanks, laboratories, hospitals, research groups, genetic testing service and individual. Since biomedical data always come with a wide variety of data type and high dimensionality, and especially genomic data are massive and fast growing, the NoSQL application databases, particularly MongoDB, with high flexibility, good scalability and low cost are attracting research groups and companies [12]. Recently, mongodb reported that a UK government owned company Genomics England is using MongoDB. The flagship project running by Genomics England is sequencing 100,000 whole genomes from patients with rare diseases and their families as well as patients with common cancers [13].

However, without strong security support and careful setting, the weakness security of this kind of NoSQL database will cause serious issues. By default MongoDB has Authentication Weakness which means that password credentials are not required immediately. Secondly, It is exposed to the internet by default. Furthermore, all data is sent in the clear and can be captured in an ARP Poison attack or man in-the-middle attack. In addition, any created user could access to the whole database by read-only mode in default, it violates the principle of Least Privilege that any user has the minimum number of privileges necessary to accomplish the work done in their role in the system. Without correct configuration, thousands of DBs are being hacked. When building up and running the database comes first and security is left behind, it's only a matter of time before data breach or information loss of genomic data.

Security of cloud computing of genomic data

As the decreasing cost of NGS have led to the fast accumulation of large sequencing data sets and leveraging these data requires large-scale computational resources, cloud computing becomes a solution for storage and process of genomic data in many medical research centers and healthcare providers. For example, large collaborative genomic projects such as the related Pan-Cancer Analysis of Whole Genomes (PCAWG) effort [14], the National Cancer Institute (NCI) Cancer Genomics Cloud (CGC) Pilots are using Cloud platform for computing and data storage [15].

However, with the large scale and high sensitivity of the genomic data, the risk of cloud computing should be critically considered. As the hallmark of cloud are elasticity, scalability and globally accessible, it also introduces the cloud characteristic vulnerabilities including internet protocol

vulnerabilities, data recovery vulnerability and unauthorized access to management interface [16]. Because of resource pooling and elasticity characteristics in cloud computing, a particular resource may be assigned to one user and allocated to another user later. A malicious user can recovery and obtain the data of previous users using data recovery techniques. This kind of data recovery vulnerability can pose significant threats to sensitive data.

In addition, Encryption of large scale of genomic data in rest (stored in disk) and in flight (during transmission) have to be ensured. But performing complex analysis on encrypted genomic data is still challenge for cloud computing. Studies for confidential query and access patterns hiding from the cloud are still under intensive study. As discussed in [17], homomorphic encryption (HE) enables the process of certain computations on the encrypted data directly, and retrieving only decrypted final result, and consequently preserving the confidentiality of genomic data from a third party cloud provider. Although HE is the state-of-the-art cryptographic technique, it has limited adoption because of lack of flexibility or impaired scaling to real-size genomic datasets.

6. Action item

Protecting genomic data and mitigating the potential risks would require the actions during data generation, storage, transmission, analysis, data sharing. All these are cooperated by biobanks, data centers, researcher, policy maker and individual users.

For data centers, biobank and researcher

At first, sensitive genomic data should not be transferred through internet in clear text such as FTP or HTTP protocol. Encryption of data in storage and transmission is an important and first step for ensuring data security. Vulnerability scanning is a key step to identify potential risk for organizations or data centers when building their data analysis pipeline. Limited data access based on personal role and a strict least-privileged authorization policy will reduce potential risk of sensitive data leakage. Furthermore, physically separation of metadata and genomic data will help to minimize the risk of re-identification of anonymous genomic data. It is also important for researcher to keep data security in a high priority and familiar with common vulnerabilities when developing new bioinformatic tools or building databases.

For Privacy policy of genomic data

Policy makers need hard working to fill the gap between the rapid advancement of next generation sequencing and privacy protecting of sensitive data. Up to now, new policies and Regulations are adopted for protecting genomic data. As an example, the public genetic-genealogy database Ysearch.org website, which was used on the route of re-identification, is no longer accessible as a result of the EU General Data Protection Regulation (GDPR) that went into effect on May 25th 2018. Meanwhile, California Consumer Privacy Act (CCPA) with strong privacy protections was signed by California Gov. this year and will be effective on 2020.

For individuals

we should be aware of benefits and risks of sequencing our genome. As more and more people are benefited from the genomic and other "omic" studies, specially those with life-threatening disease such as cancer, at some point of our lifetime, we may have our genome sequenced. Thus, it is important for individuals to learn about how your data will be used and understand your rights as well as the organization's data policies and practices.

7. Conclusion

Fast progressing of next generation sequencing technologies is providing unprecedented opportunities to change health care practice. Determining our genomic information becomes a vital part for health assessment in precision medicine. It could be also true that you sequence and access your genomic data through your smartphone in the near future. Much more attention and effort are required considering the security and privacy challenge of such special and high sensitive genomic data.

8. References

[1] https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

[2] Ayday, E., De Cristofaro, E., Hubaux, J.-P., Tsudik, G.: Whole genome sequencing: revolutionary medicine or privacy nightmare? Computer 2, 58–66 (2015)

[3] Rabbani B, Nakaoka H, Akhondzadeh S, Tekin M, Mahdieh N. Next generation sequencing: implications in personalized medicine and pharmacogenomics. Molecular bioSystems. 2016;12:1818-30.

[4] Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature. 1990;345:458-60.

[5] Haga SB, Friedman B, Richard G. Considering the Benefits and Risks of Research Participants' Access to Sequence Data. *Genet. Test. Mol. Biomarkers* 21(12), 717-721 (2017).

[6] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321–4. doi: 10.1126/science.1229566.

[7]Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15(6):409–421. doi: 10.1038/nrg3723.

[8]Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, et al. Privacy in the Genomic Era. ACM computing surveys. 2015;48.

[9] Lindor Noralane M. Personal Autonomy in the Genomic Era. Video Proceedings of Mayo Clinic Individualizing Medicine Conference; 2012. [http://bcove.me/lm00e8z7]

[10] https://www.shodan.io

[11]Ney P, Koscher K, Organick L, Ceze L, Kohno L. Computer Security, Privacy, and DNA Sequencing: Compromising Computers with Synthesized DNA, Privacy Leaks, and More. The 2017 USENIX Security Symposium. https://dnasec.cs.washington.edu/.

[12]. Iasmini Lima, Matheus Oliveira, Diego Kieckbusch, Maristela Holanda, Maria Emília M. T. Walter, Aletéia Araújo, et al, "An evaluation of data replication for bioinformatics workflows on NoSQL systems", *Bioinformatics and Biomedicine (BIBM) 2016 IEEE International Conference on*, pp. 896-901, 2016.

[13]https://www.mongodb.com/press/genomics-england-uses-mongodb-to-power-the-data-science-behind-the-100000-genomes-project

[14] Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nature reviews Genetics. 2018;19:325.

[15] International Cancer Genes Consortium. PCAWG Data Portal and Visualizations. *ICGC* http://docs.icgc.org/pcawg/ (2017).

[16] Pericherla Satya Suryateja Threats and Vulnerabilities of Cloud Computing: A Review International journal of computer sciences and engineering 6(3) March 2018

[17] Sousa JS, Lefebvre C, Huang Z, Raisaro JL, Aguilar-Melchor C, Killijian MO, et al. Efficient and secure outsourcing of genomic data storage. BMC medical genomics. 2017;10:46.