Processing of genomic data with privacy preserving methods

Rastislav Hekel, Ondrej Pös
Vedecký park Univerzity Komenského, Bratislava

Genomic data become more available and affordable because of the advancements in sequencing technologies. Analysis and interpretation of these data, especially genomic variants, is essential for development of personalized medicine. Accordingly, a sequenced genome is likely to become the standard part of personal medical record. Nevertheless, processing and storing genomic data introduces a risk of abuse by potential adversary, since a genome contains sensitive personal traits and is the ultimate person identifier. Therefore, from security point of view, genomic data must be treated and protected like personal data. On the other hand, it is crucial to keep the data available for further research. This conflict of interests is a motivation for development of novel privacy preserving methods for storage and processing of genomic data.
Keywords: genomic data, personal data, privacy, protection

Spracovanie genomických údajov pomocou metód so zachovaním súkromia
Genomické údaje sa stávajú čoraz dostupnejšími vďaka pokrokom v sekvenčných technológiách. Analýza a interpretácia týchto údajov, hlavne genomických variantov, je kľúčová pre rozvoj personalizovanej medicíny. V súlade s tým sa sekvenovaný genóm pravdepodobne stane štandardnou súčasťou zdravotnej karty. Spracovávanie a uchovávanie genomických údajov však súčasne predstavuje riziko ich zneužitia potenciálnym protivníkom, pretože genóm obsahuje citlivé osobné údaje a je definitívnym osobným identifikátorom. Z tohto dôvodu musia byť genomické údaje, z hľadiska bezpečnosti, spravované a chránené ako osobné údaje. Na druhej strane je veľmi dôležité, aby genomické údaje boli k dispozícii pre ďalší výskum. Tento konflikt záujmov je motiváciou na vývoj nových metód pre bezpečné uchovávanie a spracovávanie genomických údajov.
Kľúčové slová: genomické údaje, osobné údaje, súkromie, ochrana

Newslab, 2019; roč. 10 (1): 28 – 31

Introduction

A biological sample collected in the beginning of genomic analysis contains a genome in chemical form as a DNA molecule. In massive parallel sequencing the molecule is fragmented and placed on sequencing platform where fragments are read in parallel creating digital sequences of DNA bases called reads[1]. These reads are randomly ordered with unknown direction and unknown DNA strand of origin. If genomic reads belong to an organism with known reference genome, such as human, they are sorted in process called mapping.

Aim of the mapping is to reconstruct the original genome sequence. Each read is aligned and thus mapped to the most probable region of origin on the reference genome. An aligned and mapped read is often simply called alignment. Reads without a sufficiently probable match are considered unmapped. It is worth to mention that mapping process requires significant computer power, hence running it on computing server in parallel is reasonable. Set of aligned reads is de facto a digital copy of the DNA contained within the biological sample[2].

Aligned reads reveal differences between a sequenced and the reference genome, called genomic variants. Whole set or even specific selection of genomic variants is unique to each individual hence a genome is the ultimate person identifier. Moreover, some of these variants reveal sensitive information such as predisposition to diseases (e.g. cancer) or physical traits (e.g. race, eye colour)[3]. ***(Figure 1)***

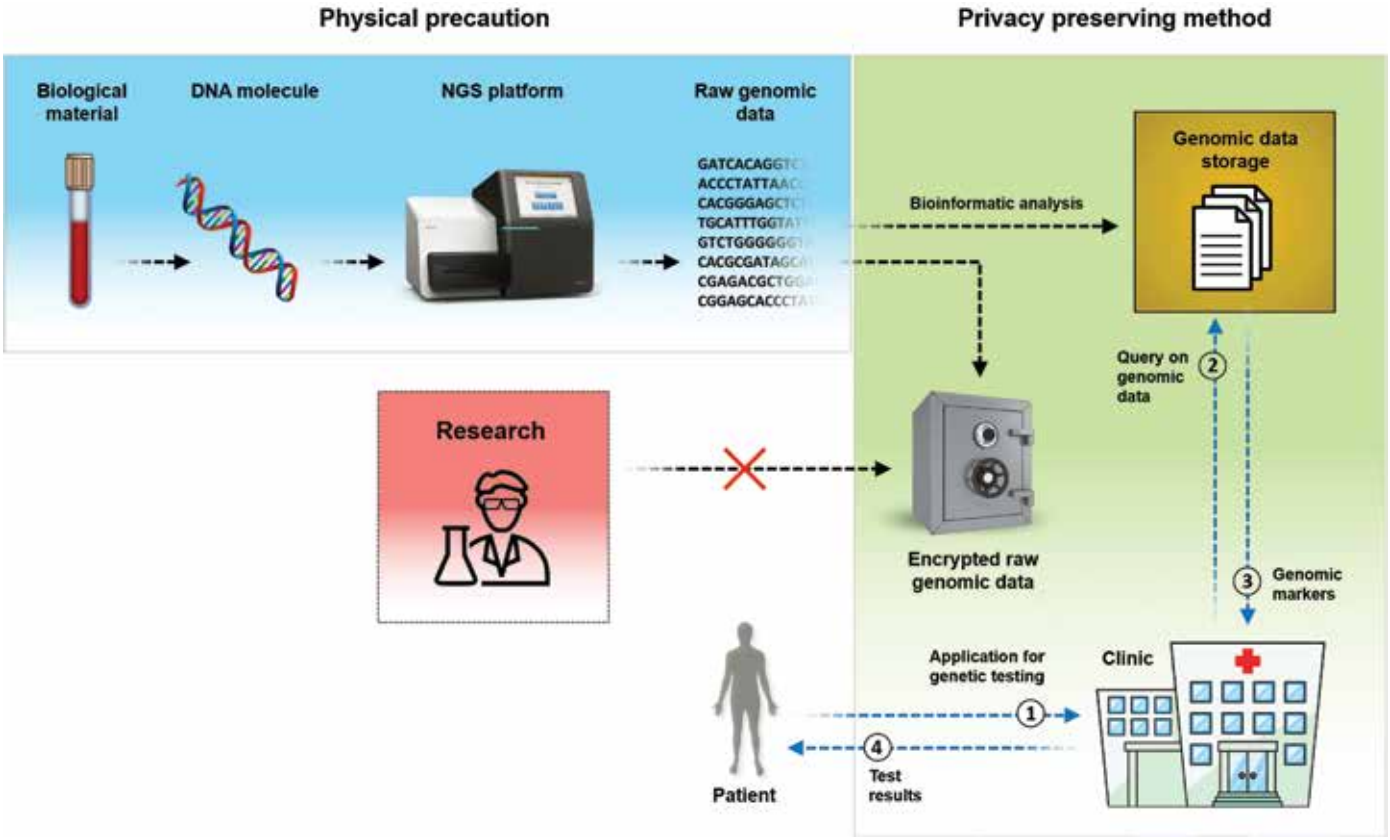
Typical clinical genetic test is a query on a personal genomic data. Genetic test can be conducted with goal to select suitable treatment or to evaluate risk for a particular disease. Usually, the outcome of this test is based on presence of specific variants associated with a tested trait. This provides a great incentive to study genomic variants, which are de facto the backbone of personalized medicine. Linking variants to specific traits is a subject of variant associations studies. Genome-Wide Association Study (GWAS) is common type of variant study in which genomes of many participants with varying phenotypes are compared for a particular trait or a disease. If one variant is more common in the group with an observed trait, the variant is said to be associated with it[4].

Most variant studies require only limited access to short regions of specific genes, such as diagnosis of specific disorder with a known set of causal genes. In addition, some analyses do not even require information about genomic variants at all. For example, chromosome aneuploidy detection needs only information about number of reads aligned to individual chromosomes, in other words coverage, that is not major subject to misuse or person identification[5].

Although the final product of a variant analysis is not the genome itself, there is a preference to keep raw aligned reads along the interpreted variants for the following reasons:

– Algorithms for variant calling are not mature, they can have various settings and trade-offs.

Figure 1. *Common processing of genomic data intended for medical use. DNA is collected from biological material and placed on sequencing platform. Until now privacy is protected by standard physical precautions. Typically, personal variants or other medically relevant personal data are derived from a genome and the genome itself is encrypted and archived. With consent of a patient his personal genomic data can be tested for specific markers by a medical clinic. Genomic data are subject to potentional abuse, hence they need to be protected by some privacy preserving method during the whole genetic testing. Researchers have no way to utilize non-personal data within the sequenced genome without direct consent of patient and support from both the medical clinic and data storage.*



- Disease such as cancer can cause specific variations in diseased cells. These can be misclassified as sequencing errors by only looking at variant calls, while examining raw reads can reveal the true cause.
- It is impossible to know which (novel) variants are going to be proved as significant in the future.

Until the sequencing the privacy of the genome can be preserved by standard physical precautions. However, the sequencing creates digital copy of a genome and from there it must be secured digitally to prevent unwanted copying, modifying and sharing. The interpretation of genomic data intended for treatment can be abused in hands of an potential adversary[6]. In contrary, it is crucial to keep genomic data available for further research[7].

Privacy problems

In the literature, there are several solutions suggested for privacy and security of digital health records based on de-identification and aggregation methods. However, these solutions are not applicable to personal genomic data as genome itself is an ultimate identifier of an individual. Akgün et al. 2015 provide great overview on privacy processing of genomic data and state the following major problems which need to be solved[8]:

- **Private read alignment on public cloud**
  Alice wants to make a sequence alignment for her whole genome on a public cloud controlled by Bob, without revealing the genome to Bob.
- **Query on private genomic data**
  Alice wants to test her genome for some biological trait. The test is provided by Bob, who must query Alice's genome with publicly known markers for that trait. Alice does not want to reveal her whole genome to Bob.
- **Query on private genomic database**
  Alice want to test a hypothesis using a genomic database, while Bob (responsible for the database) wants to preserve the privacy of the data-owners.
- **Privacy-preserving sharing of private statistical database**
  GWAS produces population statistics for associations between variants and specific traits. Alice wants to query statistics relevant for her study, while Bob (responsible for GWAS) do not want to reveal if some individual is part of GWAS.

Cryptographic solutions

1. Secure Multiparty Computation (SMC)

Sequence alignment is likely to be outsourced to public clouds due to high computation cost. Public cloud is

considered as an insecure environment where private data can be obtained by an adversary. Due to this concern secure computation scheme must be used instead of standard alignment algorithms. SMC is basis of some proposed solutions[9]. The method allows two or more entities jointly compute on private data without revealing the data to each other or a third party. This enables outsourcing most of computation intensive read mapping without disclosing genetic information. Moreover, work of Jagadeesh et al. 2017 showed how SMC can be used to securely identify causative variants in individuals between multiple parties. More precisely, they focus on variants in Mendelian patients and use SMC methods based on Yao's protocol. The computation can be run on the whole genomes provided by various parties (e.g. institutions, patients) to jointly discover the causative variants, while not revealing the genomes to each other[10].

2. Homomorphic Encryption

Variant association studies require unrestricted access to large databases of genotype and phenotype data to compute reliable statistics. These data collected from volunteering patients are at risk of privacy breach when stored in unencrypted form. Kim et al. 2015 proposed solution to this problem: encrypting both genotype and phenotype data by homomorphic encryption scheme. They take several statistical algorithms commonly used in GWAS studies and altering them so they can be run on the encrypted data. Homomorphic encryption is allowing to directly compute on the encrypted data without knowing a passphrase[11]. Work of Sousa et al. 2017 enables a user to securely store and retrieve millions of genomic variants of all types for one or multiple individuals on the cloud. Variants are encrypted with symmetric key and can be efficiently searched without revealing anything to the cloud provider. They use homomorphic encryption and private information retrieval techniques[12]. Novel approach proposed by Shimizu et al. 2016 combines efficient string data structure called positional Burrows–Wheeler transform (PBWT) with two cryptographic techniques called additive homomorphic encryption and oblivious transfer[13].

3. Differential privacy

It is impossible to publish information from a private statistical database without revealing some amount of private information, therefore a small number of database queries can reveal a presence of some record. For instance, the presence of specific genome with known minor allele freqencies can be detected in some statistical dataset. First, these frequencies must be compared against reference population frequencies and then against frequencies from the dataset. Finally, the difference is evaluated with statistical t-test.[14]. Differential privacy solves this problem by maximizing the accuracy of queried statistics while minimizing the information about presence of specific record. Solution offers trade-off between utility (accuracy) and privacy. Unencrypted data is available only through special queries which add noise to the result of each query. Differentially private genomic databases differing from each other by only one individual's data, have indistinguishable statistical features.

Processing of raw genomic data

Clinical genetic testing is becoming common in personalized medicine and each conducted test introduces a risk to genomic privacy of a patient. Since most of these tests are based on presence of specific genomic variants, the typical solution is to extract all the variants from an underlying genomic data and store them in encrypted form suitable for secure analysis[11,12,15]. The raw genomic data, typically in form of aligned reads, are not further considered. They are usually encrypted and stored separately so they can be reused in the future when novel variants are discovered.

Only one of all examined works on the subject of genomic privacy is focused at secure storage, retrieval and processing of a raw genomic data. Ayday et al. 2014 propose a scheme that stores genomes of patients in form of encrypted alignments, in a public *biobank*. A *medical unit* can request a genomic region from a *biobank* without revealing the scope of the request, so the *biobank* can not infer the nature of the genetic test behind the request. The *biobank* provides only alignments that include at least one base from the requested range. Nucleotide bases outside of this range are masked from the *medical unit* while no decryption is involved. Returned alignments are decrypted in requested range at the *medical unit*. Encryption keys are stored separately at *masking and key manager*, because not all patients are capable of protecting their keys on a private device. Furthermore, when patient controls his key, he must be involved in all operations of medical unit, related to his genome, which is not practical when conducting research. Identities of *medical units* or patients are not revealed to the *masking and key manager*[15].

Protection from third party

Even though conventional digital security methods can protect against unauthorized use of personal data, they do not protect the data from misuse by an authorized third party. The data can be potentially accessed by a third party in legal way under data sharing policy without the owner's knowledge. According to some opinions, this problem cannot be efficiently solved using current technological solutions. For this reason, the solution seems to remain on legislation and on the awareness of the society[16]. Although this may be true, the dynamic consent approach give the owner an ability to retain control over his personal data. Instead of giving single consent to share the data under some policy at the time of sampling, the owner can choose to share his data with trusted party on demand. This aspect is considered to be a core element of modern information privacy[17].

Conclusion

All of the reviewed schemes for privacy preserving processing of genomic data do not consider or completely encrypt the underlying raw genomic data. Retrieval, decryption or interpretation of the raw genomic data is available only through special procedures by authorized parties. Besides, some sort of consent is required when requesting the data. As a result, information produced by DNA sequencing is constrained or unavailable for further research by scientific

community. Nevertheless, information within raw genomic data not related to personal variants should be fully available for scientific studies. This information is not considered private or individual specific, therefore it can be utilized by genomic studies unrelated to variants. With this intention a novel method for secure and reversible masking of personal variants would help both patients and scientific research.

REFERENCES

1. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456: 53-59.
2. Bleidorn C. Alignment and Mapping. Phylogenomics. Springer, Cham; 2017. pp. 105-125.
3. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015; 17: 405-424.
4. Genome-Wide Association Studies Fact Sheet. In: National Human Genome Research Institute (NHGRI) [Internet]. [cited 7 Jan 2019]. Available: https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/
5. Fiorentino F, Biricik A, Bono S, et al. Development and validation of a next-generation sequencing-based protocol for 24-chromosome aneuploidy screening of embryos. Fertil Steril 2014; 101: 1375-1382.
6. Frizzo-Barker J, Chow-White PA, Charters A, Ha D. Genomic Big Data and Privacy: Challenges and Opportunities for Precision Medicine. Comput Support Coop Work. Springer Netherlands 2016; 25: 115-136.
7. Shen H, Ma J. Privacy Challenges of Genomic Big Data. Adv Exp Med Biol 2017; 1028: 139-148.
8. Akgün M, Bayrak AO, Ozer B, Sağıroğlu MŞ. Privacy preserving processing of genomic data: A survey. J Biomed Inform 2015; 56: 103-111.
9. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014; 15: 409-421.
10. Jagadeesh KA, Wu DJ, Birgmeier JA, et al. Deriving genomic diagnoses without revealing patient genomes. Science 2017; 357: 692-695.
11. Kim M, Lauter K. Private genome analysis through homomorphic encryption. BMC Med Inform Decis Mak 2015; 15(Suppl 5): S3.
12. Sousa JS, Lefebvre C, Huang Z, et al. Efficient and secure outsourcing of genomic data storage. BMC Med Genomics 2017; 10: 46.
13. Shimizu K, Nuida K, Rätsch G. Efficient privacy-preserving string search and an application in genomics. Bioinformatics 2016; 32: 1652-1661.
14. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008; 4: e1000167.
15. Ayday E, De Cristofaro E, Hubaux J-P, Tsudik G. The Chills and Thrills of Whole Genome Sequencing [Internet]. 2013. Available: http://arxiv.org/abs/1306.1264
16. Savage N. Privacy: The myth of anonymity. Nature 2016; 537: S70-S72.
17. Erlich Y, Williams JB, Glazer D, et al. Redefining genomic privacy: trust and empowerment. PLoS Biol 2014; 12: e1001983.

Mgr. Rastislav Hekel
Vedecký park Univerzity Komenského
Ilkovičova 8, 841 04 Bratislava
e-mail: rastislav.hekel@geneton.sk