# Varlock: privacy preserving storage and dissemination of sequenced genomic data

Rastislav Hekel[1,2,3], Jaroslav Budiš[1,3,4], Marcel Kucharík[1,4], Jan Radvanszky[1,4,5], Tomáš Szemes[1,2,4]

1. Geneton Ltd., Bratislava, Slovakia
2. Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia
3. Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia
4. Comenius University Science Park, Bratislava, Slovakia
5. Institute of Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava, Slovakia

## Abstract

**Introduction:** Current and future applications of genomic data may raise ethical and privacy concerns. Processing and storing genomic data introduces a risk of abuse by a potential adversary since the human genome contains information about sensitive personal traits. For this reason, we developed a privacy preserving method, called Varlock, for secure storage and dissemination of sequenced genomic data.

**Materials and methods:** The Varlock uses a set of population allele frequencies to mask personal alleles detected in genomic reads. Each detected allele is replaced by a randomly selected population allele concerning its frequency. Masked alleles are preserved in an encrypted confidential file that can be shared, in whole or in part, using public-key cryptography.

**Results:** Our method masked personal variants and introduced new variants called on an individual's genome, while alternative alleles with lower population frequency were masked and introduced more often. We performed joint PCA analysis of personal and masked VCFs, showing that the VCFs between the two groups can not be trivially mapped. Moreover, the method is reversible; therefore, personal alleles can be unmasked in specific genomic regions on demand.

**Conclusion:** Our method masks personal alleles within mapped reads while preserving valuable non-sensitive properties of sequenced DNA fragments for further research. Accordingly, masked reads can be stored publicly, since they are deprived of sensitive personal information. Personal alleles may be restored in arbitrary genomic regions for interested parties: patients, medical units, and researchers.

**Keywords:** genome, privacy, personal data

# Introduction

The ongoing advancements in DNA sequencing technologies drive the increasingly complex and accurate interpretation of genomic data, together with the development of precision medicine [1]. A potential adversary can abuse genomic data since they carry sensitive personal information, such as disease risks and phenotypic traits [2]. Accordingly, genomic data are regulated as personal data [3]; nevertheless, keeping the data open for further research is essential [4].

In general, many genomic analyses stand on the presence of short genomic variants; hence the typical solution of the prior art is to extract these variants from the underlying genomic reads [5–7]. The prior art stores these variants in a secure form, while discarding the genomic reads, or encrypting them completely, so they can be reanalysed in future. However, manual examination of reported variants in aligned genomic reads is a common practice to confirm a finding, and specific variations can be missing from variant calls due to their misclassification as sequencing errors [8]. Moreover, current variant calling algorithms are not mature, and it is unknown which type of data produced by the sequencing process will be necessary for future algorithms [9]. Additionally, aligned reads can be employed directly in the detection of structural variations such as copy number variations (CNVs) or aneuploidies in clinical non-invasive prenatal testing (NIPT) [10–12]. These detection methods are not dependent on short variant analyses, since they use coverage data, determined by read alignment only.

We developed the tool Varlock with two main goals. First, to keep sequenced data available without compromising the privacy of a patient, and second, to create an access control mechanism for extracting sensitive private information contained within the sequenced data. More specifically, the Varlock masks personal alleles within aligned reads of a sequenced genome, while preserving existing alignment data (coverage, quality, etc.). The method is reversible, allowing the user with access to masked personal alleles to unmask them within an arbitrary region of the associated genome. The user can also share access to a subset of the masked alleles, for example from a particular gene, with another user.

# Materials and Methods

The Varlock provides methods for masking, unmasking, and dissemination of personal alleles found in mapped reads stored in a BAM file. More specifically, the masking method

(Figure 1) masks personal alleles found in mapped reads using publicly known population allele frequencies. The output set of masked alleles represents all differences between original and masked mapped reads. The masked alleles are encrypted as a single file using an asymmetric encryption scheme (Supplement 3), so only the owner of the associated private key can decrypt them.

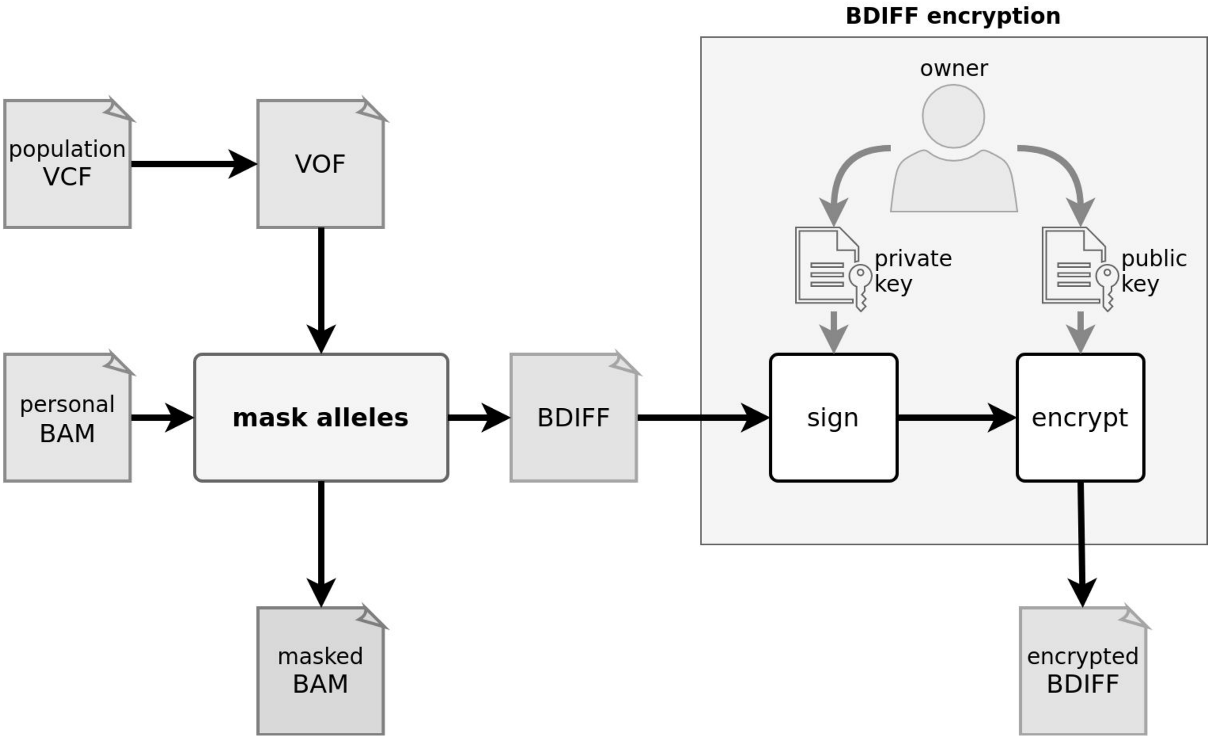

***Figure 1:*** *Workflow of the masking method, where BAM file and VOF file are processed into masked BAM and BDIFF. The BDIFF file is subsequently encrypted.*

The unmasking method (Figure 2) is a partially reversed masking method. The file with masked alleles is decrypted with the associated private key and is processed simultaneously with masked mapped reads back into personal mapped reads. The dissemination (Figure 3) method re-encrypts the file with masked alleles in an arbitrary range, making the associated subset of alleles accessible for a specific user. Firstly, the file with masked alleles is decrypted by the associated private key; secondly, a subset of masked alleles is selected, and lastly, the selected masked alleles are encrypted as a new file with the public key of a specific user.

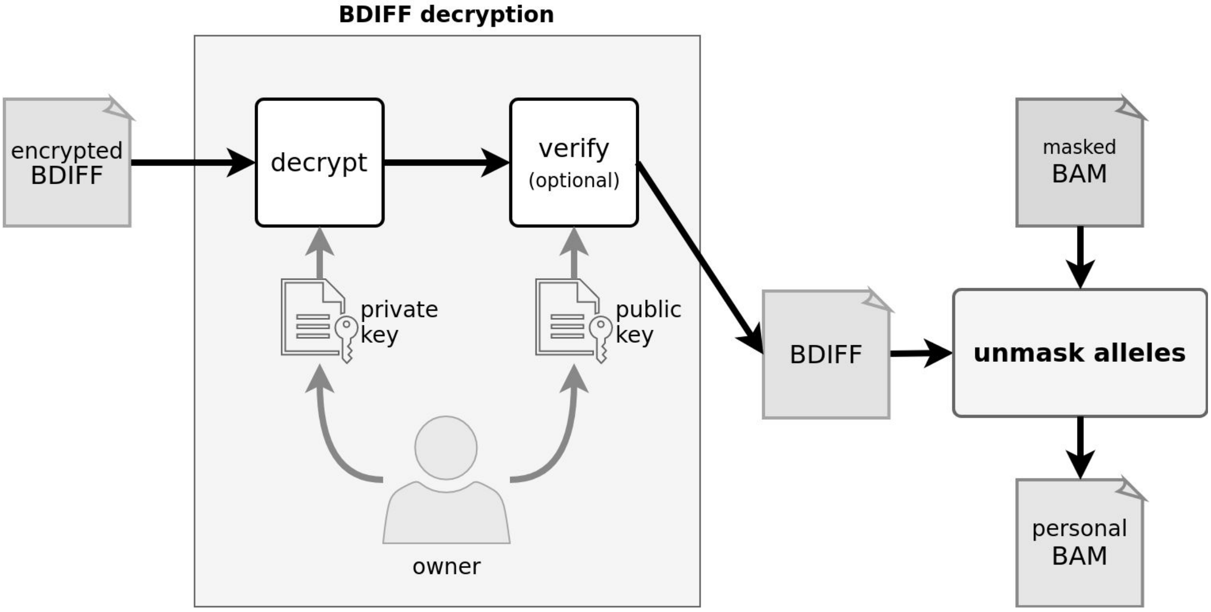

***Figure 2:*** *Workflow of the unmasking method, where a BDIFF file is decrypted and used to unmask a masked BAM file to restore a personal BAM file.*

We introduce two file formats VOF (Supplement 2) and BDIFF (Supplement 3); VOF describes population allele frequencies, and BDIFF is the format of masked alleles which is used to unmask masked mapped reads.

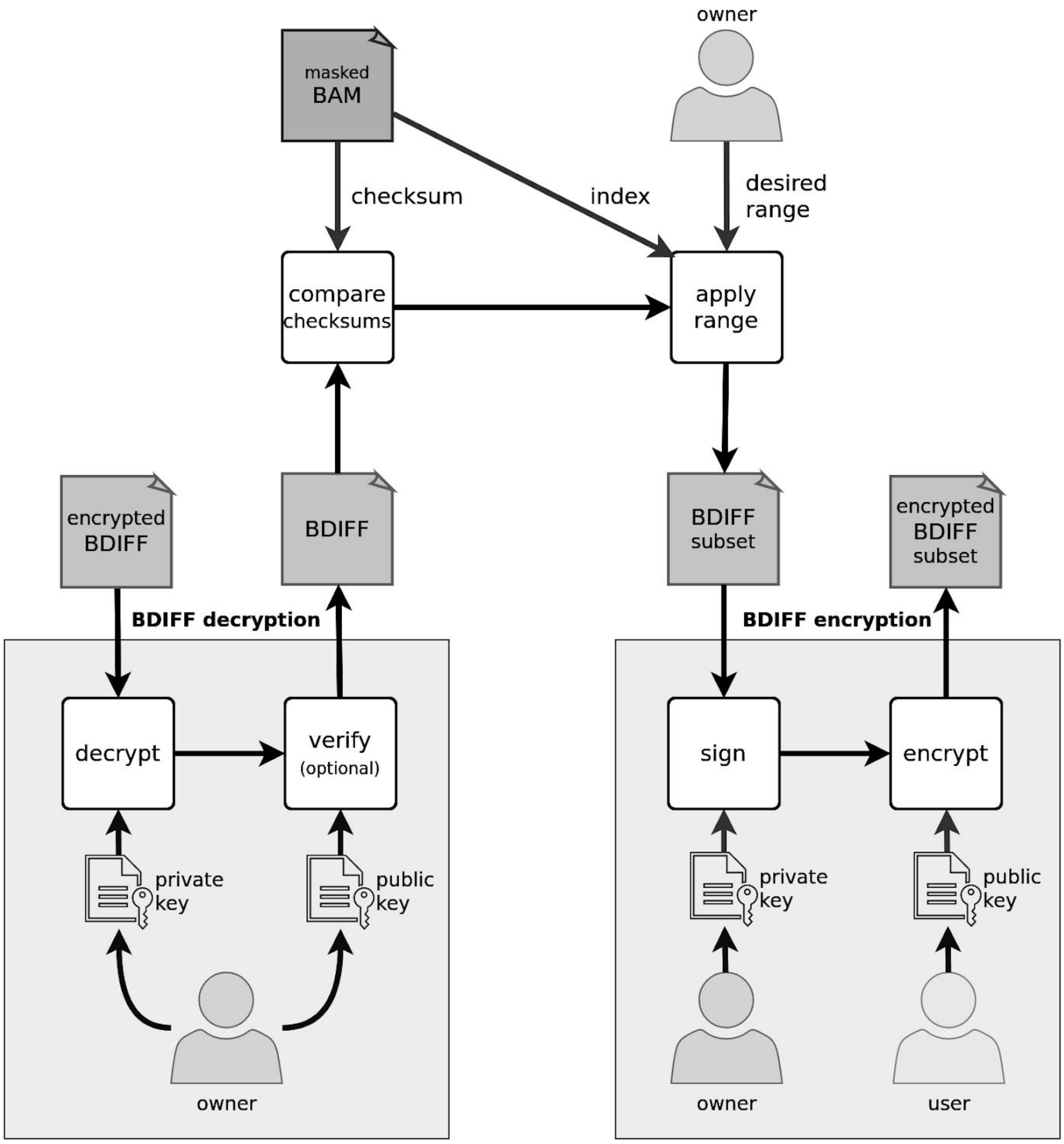

***Figure 3:*** *Workflow of a dissemination method showing decryption of BDIFF and encryption of its subrange intended for a specific user.*

## Masking of alleles

A sequenced genomic position is typically covered by multiple alignments, which may carry different alleles due to heterozygosity, sequencing, or alignment errors. Both personal alleles are equally likely to be represented in the alignments, albeit their mutual ratio can substantially vary for a given position. Therefore, each genomic position with a population variant is described by a list of alleles, and the personal pair of alleles is determined as the two most represented ones. In detail, an allele is considered personal if it constitutes at least

20% (arbitrarily chosen threshold) of alignments covering the variant. If only one such allele exists, the position is evaluated as homozygous, and two identical alleles are assigned to the position. If two alleles with a sufficient representation exist, the position is considered heterozygous, and two different alleles are assigned to the position. If there are more than two sufficiently represented alleles, the variant position is skipped by the method.

The process of masking and unmasking alleles per given position has several steps (Figure 4). The population allele frequencies defined in VOF are multiplied with each other to produce a probability matrix of every possible pair of alleles at a given genomic position (Supplementary Methods, Section 4). The pair of masking alleles is drawn randomly from this probability matrix as a replacement for the pair of personal alleles assigned previously. If a reference allele replaces both personal alternative alleles, a variant can not be detected in masked mapped reads; therefore, it is masked. Conversely, if an alternative allele replaces either of the personal reference alleles, a new variant can be called at this position in masked mapped reads; thus, it is introduced. All personal alleles within the alignments covering a variant are replaced by masking alleles. However, personal alleles may be replaced by the same pair of masking alleles, which is the most common case. Remaining alleles found within the alignments are considered to be sequencing or alignment errors and are not replaced or replaced by other than masking alleles.

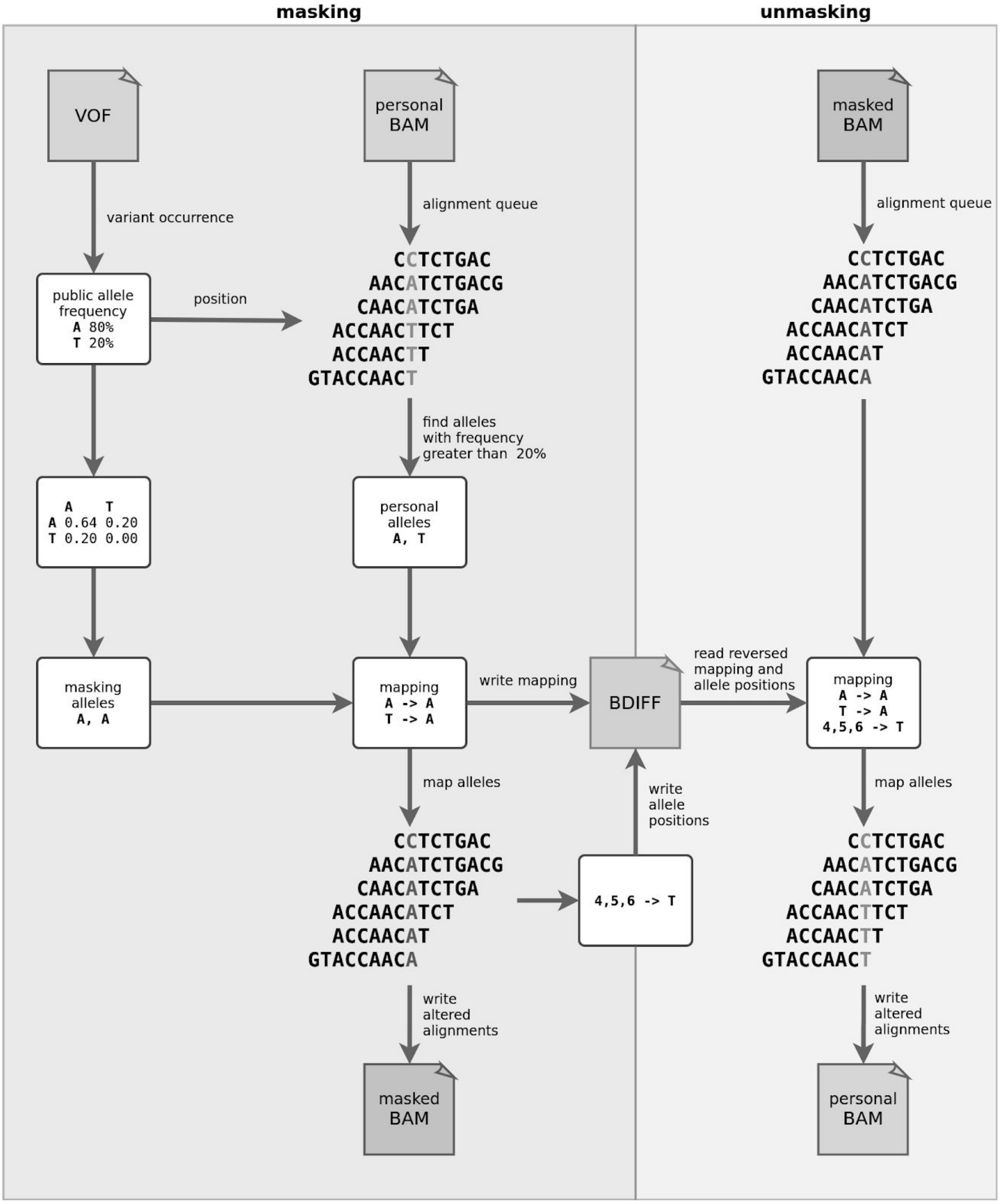

***Figure 4:*** *Flow of masking and unmasking alleles at a single variant position within covering alignments. The masking is represented as “mask alleles” in Figure 1, and the unmasking is represented as “unmask alleles” in Figure 2.*

Both personal and masking pair can be either homozygous or heterozygous, leading to one of the following cases:

**Homozygous to homozygous:** Two identical masking alleles replace two identical personal alleles. Most often, two reference alleles are replaced by the same two alleles, since reference allele is typically the most common one in both personal mapped reads and population allele frequencies. If pairs are identical, no actual masking occurs. *Possible outcome: masked variant, introduced variant, replaced variant, none.*

**Heterozygous to homozygous:** Two identical masking alleles replace two different personal alleles. Reference and alternative alleles are often replaced by two reference alleles, which results in masking of a personal variant. *Possible outcome: masked variant, replaced variant.*

**Homozygous to heterozygous:** Two different masking alleles replace two identical personal alleles. If an alternative allele replaces either reference allele, a new variant emerges. Possible *outcome**:** introduced variant, replaced variant.*

**Heterozygous to heterozygous:** Two different masking alleles replace another two different personal alleles. Personal and masking pairs of alleles are often identical, so no actual masking occurs. If only one personal allele is identical to a masking allele, the other personal allele is masked with the remaining masking allele. In this case, the variant can not be masked since the alternative allele can be replaced only by another alternative allele. *Possible outcome**:** replaced variant, none.*

## Unmasking of alleles

All alleles within masked mapped reads, or their specific subset, can be unmasked by BDIFF file, containing replaced personal alleles and deleted qualities. This operation transforms masked mapped reads to personal mapped reads. User has to provide masked mapped reads and an associated encrypted BDIFF file along with the RSA private key whose public counterpart was used in the BDIFF encryption. The decryption of unmapped reads is handled separately, and the user can choose whether to decrypt them.

The first step of unmasking method is the decryption of the encrypted BDIFF file (Supplement 3.1). The algorithm reads the encrypted AES key and the file signature from the start of the file. The AES key is decrypted with a provided private key and then used to decrypt an actual encrypted BDIFF file. The decrypted file is verified with a public key against its signature to prove its origin.

## Dissemination of alleles

A holder of the private key that was used to encrypt a BDIFF file can disseminate alleles described by BDIFF file and associated masked mapped reads by re-encrypting the BDIFF file in desired genomic range. A BDIFF file is first decrypted by the private key and then encrypted by a public key of another user who can decrypt the file later. If a subrange of effective range for re-encryption is provided, only records inside or intersecting this range are considered, and this range becomes the effective range of the new BDIFF file. The re-encryption process can be repeated with different combinations of genomic ranges and public keys, producing different accesses for individual users. In addition, the decrypted BDIFF file can be verified with the holder's public key by comparing the checksum of masked mapped reads to the checksum stored in the encrypted BDIFF file header. This ensures that the BDIFF file belongs to the masked mapped reads and that they were not modified.

## Validation

To validate the Varlock, we collected a set of 37 clinical exomes from the central European population. The DNA samples were sequenced on Illumina platform following enrichment and library preparation using TruSight One clinical exome sequencing panel according to the manufacturer's instructions. Next, we called variants on each exome with a fine-tuned variant calling pipeline comprising BWA-MEM mapper (Li and Durbin 2009) and DeepVariant caller (Poplin et al. 2018), producing 37 BAM files and the same number of corresponding VCF files. Finally, we masked each BAM file with the Varlock and called variants on these masked BAM files subsequently, producing the same number of VCF files.

As the source of population variants, we used the Genome Aggregation Database version 3 (gnomAD v3) mapped to GRCh38 reference [13], which spans 71,702 genomes from unrelated individuals of various ethnicities. We downloaded the database in the form of a single VCF file, selected passing single nucleotide variants within ranges of Trusight One clinical exome panel, and merged duplicate variant positions as multiallelic. Finally, we converted the file to VOF format intended for masking.

We performed two separate validation analyses: (1) the single case study on a selected sample and (2) the PCA masking analysis on the whole set of samples. In the single case study, we used non-Finnish European gnomAD population variants in VOF format to best match the central European population of the sample. The output files of the masking method - BDIFF and masked BAM were used as the unmasking method input files. In the

PCA masking analysis, we merged all the passing single nucleotide variants from both personal and masked VCFs, 74 in total, into a single VCF. The PCA analysis was performed on this file by the tool PLINK [14] twice, each time with a different VOF file. Firstly, with all gnomAD populations, and secondly, with the non-Finnish European population, since it best matches the central European population of sequenced individuals.

# Results

## Single case study

The performance of the masking method was evaluated by comparison between called variants on a single personal BAM file, called variants on the corresponding masked BAM file, and the variants from non-Finnish European gnomAD population. We selected only passing variants with total coverage and quality above 30 from both personal and masked VCF files to provide confident results. We identified five categories of variant positions from personal VCF, masked VCF, and population VCF (Figure 5). (1) *Not found*: Vast majority of variant positions in the population VCF is not found in the personal VCF. This is expected as the population VCF is called on thousands of personal genomes, and masking of rare variants tends to result in a homozygous reference. (2) *Masked*: This case occurs when a homozygous reference allele masks a homozygous alternative allele. (3) *Not masked*: Alternative allele at this position was either preserved or replaced by another alternative allele while zygosity may be changed. (4) *Introduced*: When an alternative allele replaces reference allele at a homozygous position, a new variant appears. (5) *Not covered*: Set of personal variant positions not covered by the population VCF. These are presumably rare variants or variants specific for a particular local population that was not part of the gnomAD database.

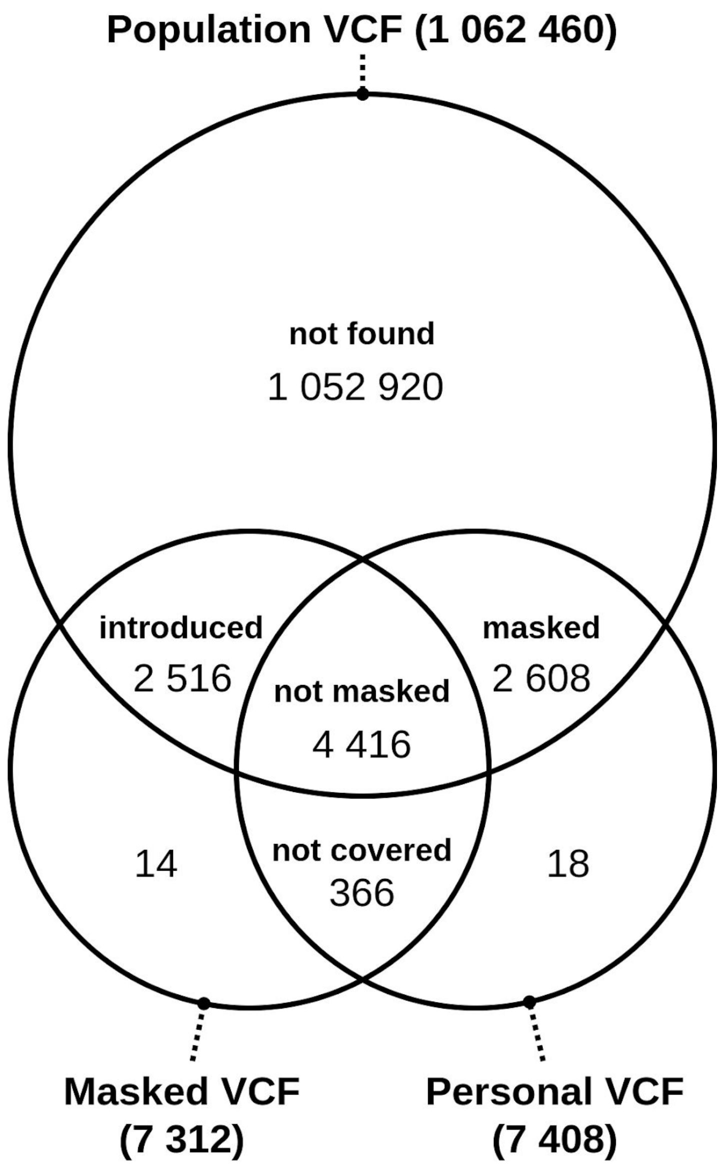

***Figure 5:*** *Intersections between sets of positions with alternative alleles from three VCF files: population VCF, personal VCF, and masked VCF.*

We compared distributions of alternative allele frequencies by VCF to show their nature and the effect of masking (Figure 6). The population VCF contains a vast amount of low frequency alleles which have a little chance to be introduced by the masking process into the masked VCF despite every variant covered by personal BAM is considered. In case of the personal VCF, personal allele frequency has expected ratio of 0.5 for a heterozygote and 1.0 for a homozygote but actual ratios may quite vary due to low coverage or sequencing errors. As can be seen, masked VCF preserves the distribution of personal allele frequency to a considerable extent.

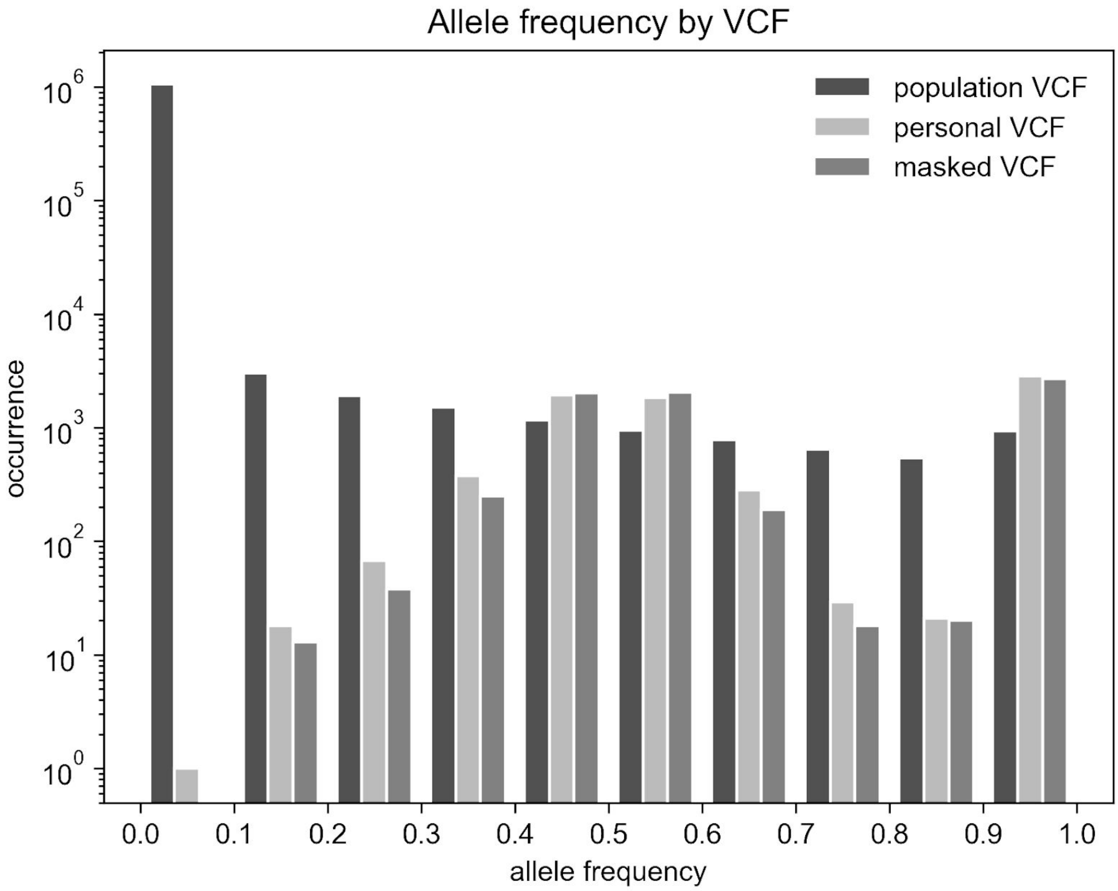

***Figure 6:*** *The distribution of alternative allele frequency reported by population VCF, personal VCF, and masked VCF.*

Furthermore, we compared the distribution of alternative population allele frequencies between the masked VCF and the not masked VCF (Figure 7). The ratio of masked alleles increases with decreasing frequency of an allele; therefore, rare variants have a higher chance to be masked by the method. Similarly, the ratio of introduced alleles increases with a decreasing frequency of an allele. On the other hand, common population alleles have a lower chance to be masked or introduced; nonetheless, they are specific for the population and not for an individual.

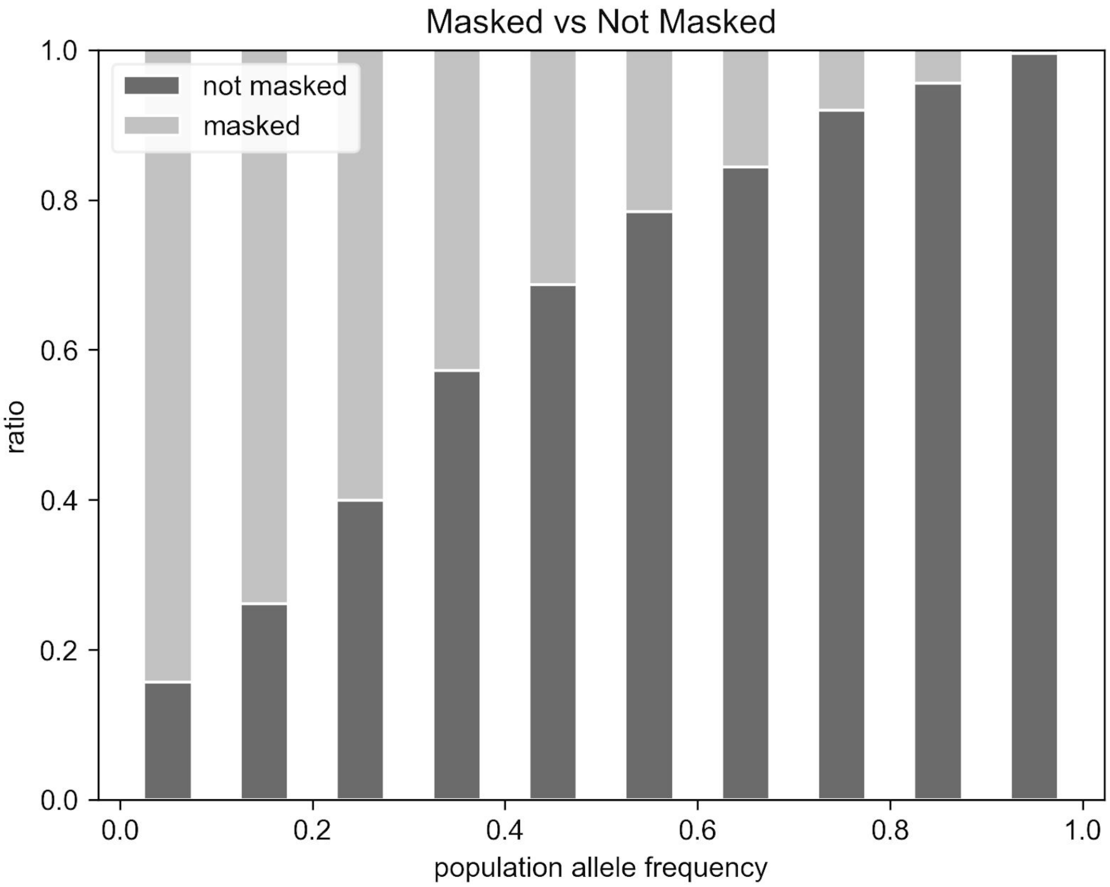

***Figure 7:*** *The ratio of masked alleles to not masked alleles and its relation to population allele frequency.*

Finally, we compared alleles in the not masked set between the personal VCF and the masked VCF. The alternative alleles from both VCFs were joined by their positions, allowing direct comparison of an alternative allele and its frequency between the two files. An alternative allele was replaced by another alternative allele in only 13 (0.29%) from a total of 4416 reported positions; thus this case is negligible. We compared frequencies of 4406 remaining positions with matching alleles between the personal and masked file and found a mismatch in 1463 (33.23%) of them. The changes of frequencies of alternative alleles in these positions were caused by the change of homozygous pair of alleles to heterozygous pair or vice-versa by the masking method.

## PCA masking analysis

We plotted the first two principal components and distinguished original and masked VCFs with a marker type. In the first case (Figure 8) the masked VCFs are clearly separated from the personal VCFs as two different groups, implying that masking using whole gnomAD variation caused a shift from the population of origin to the mixture of gnomAD populations. In the second case (Figure 9, 10), the masked VCFs can not be unambiguously mapped to corresponding personal VCFs since they stay within the same population space. Moreover, outliers - VCFs with specific genotypes are shifted into the same population cluster.

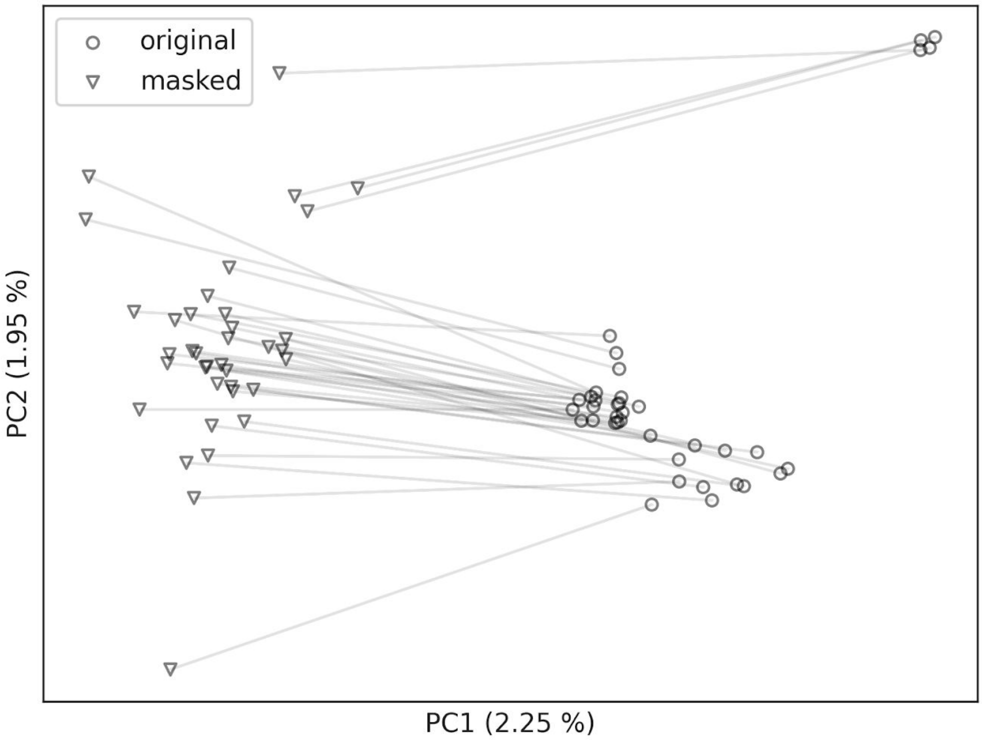

***Figure 8:*** *Personal VCFs are clearly shifted from the original local population (non-Finnish European) to VCFs masked with alleles from all gnomAD populations. Lines link individual original BAMs (circles) with their masked counterparts (triangles).*

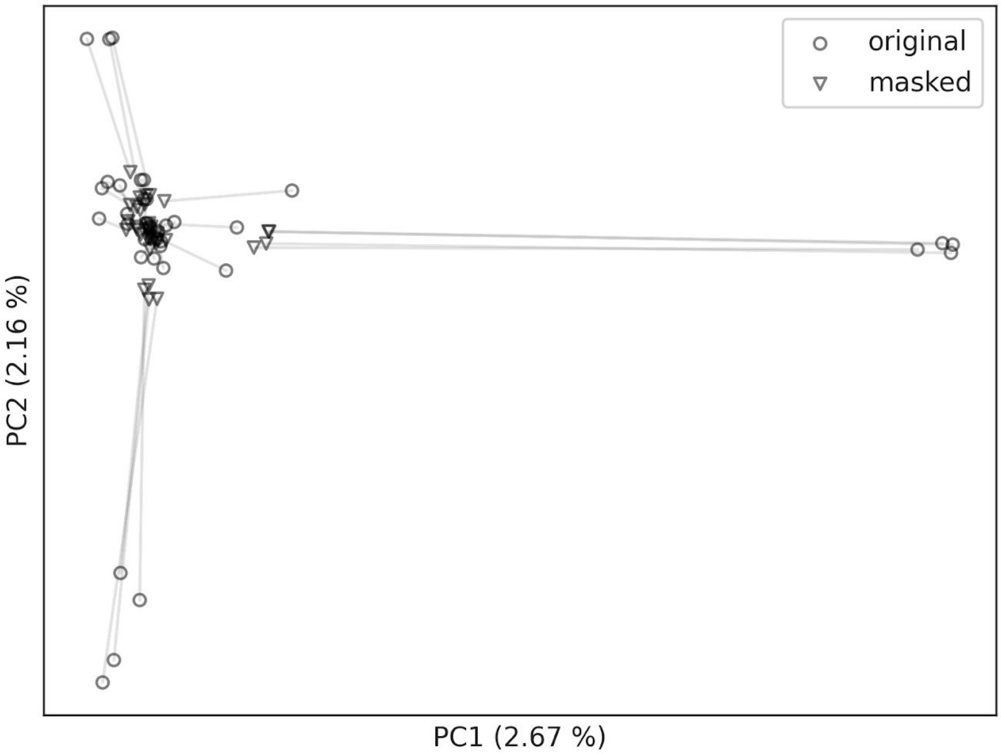

***Figure 9:*** *All masked VCFs, including outliers in their personal form, are clustered in the same region. The lines link individual original BAMs (circles) with their masked counterparts (triangles). For detail of the cluster, see Figure 10.*

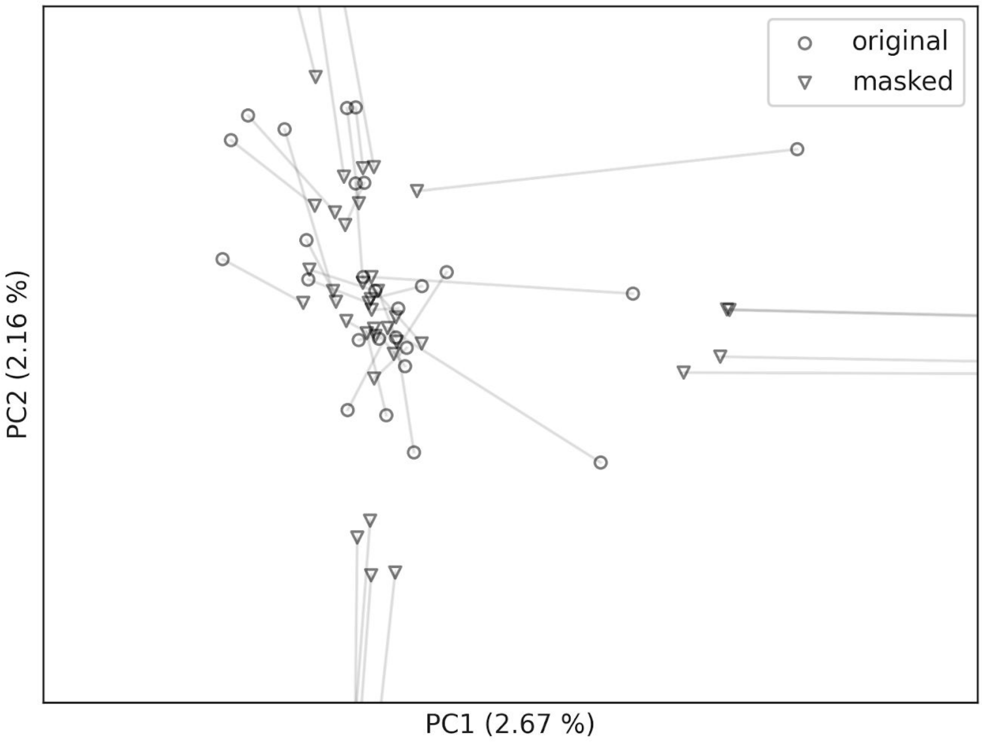

**Figure 10**

***Figure 10:*** *The detail of the cluster from Figure 9. The lines link individual original BAMs (circles) with their masked counterparts (triangles).*

# Discussion

The continuously improving interpretation of genomic data makes them prone to abuse by a potential adversary [2]; therefore, it is essential to prevent their unwanted copying, modifying, and sharing. On the other hand, open genomic data are invaluable for further research and their application in clinical practice [4]. Given these points, a practical solution to genomic privacy is a certain trade-off between privacy and utility.

The examined methods for preserving genomic privacy encrypt genomic data entirely aiming to secure personal variants [6,7,15–17]. The retrieval, decryption, and interpretation of encrypted data are available only through special procedures by authorized parties; besides, some sort of consent is required when requesting the data. As a result, access to complete genomic information produced by sequencing is restricted or limited, even for the scientific community. On the contrary, our method allows an individual to retain full control over his digital genome, supporting dynamic consent approach to access a subset of his alleles, for a clinical purpose or a scientific study.

The Varlock masks short alternative alleles within a sequenced genome and securely preserves them, allowing open access to the masked genome, which can be employed in studies unrelated to short genomic variations. In addition, we assume that a masked BAM file can not be identified as masked, since it preserves the natural distribution of alternative allele frequencies, giving an advantage against a potential adversary. However, an adversary can tell which genomic positions may be masked, given masking population allele frequencies are public, and focus on the positions that are not covered. Consequently, he could find and exploit rare personal variants. This could be mitigated by using a more robust set of masking population allele frequencies or by random masking (Supplement 4.4). In future work, we consider masking also called variants to solve this problem.

While our approach masks sensitive personal information, the genome still carries unique information, and so person re-identification by the masked genome is still possible. Although, masking more types of genomic variation, such as short tandem repeats, could make the re-identification harder. Regardless, the goal of Varlock is not to anonymize a genome, merely hide sensitive personal information. We believe that concepts behind the Varlock will find application in future medical or laboratory information management systems.

# Declarations of Interest

The authors are employees of Geneton Ltd. who participated in the development of submitted patent: *A computer implemented method for privacy preserving storage of raw genome data based on population variants* - PCT/EP2019/067336.

# Acknowledgements

This article was created with the support of the OP Integrated Infrastructure for the project: *Long term strategic research and development focused on the occurrence of Lynch syndrome in the Slovak population and possibilities of prevention of tumors associated with this syndrome*, ITMS: 313011V578, co-financed by the European Regional Development Fund.

The article was also created with the support of the OP Integrated Infrastructure for the project: *Introduction of an innovative test for screening and monitoring of cancer patients -*

*GenoScan* LBquant , ITMS: NFP313010Q927, co-financed by the European Regional Development Fund.

# Resources

The source code is available on GitHub: https://github.com/rtcz/varlock. Individual methods can be run as tests. The clinical exomes dataset used to evaluate the Varlock is not publicly available due to personal data protection but is available from the corresponding author on a reasonable request.

# References

1. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17: 507–522.

2. 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. 2016;25: 115–136.

3. Sariyar M, Suhr S, Schlünder I. How Sensitive Is Genetic Data? Biopreserv Biobank. 2017;15: 494–501.

4. Shen H, Ma J. Privacy Challenges of Genomic Big Data. Adv Exp Med Biol. 2017;1028: 139–148.

5. Ayday E, De Cristofaro E, Hubaux J-P, Tsudik G. The Chills and Thrills of Whole Genome Sequencing. Computer. 2013. pp. 1–1. doi:10.1109/mc.2013.333

6. Sousa JS, Lefebvre C, Huang Z, Raisaro JL, Aguilar-Melchor C, Killijian M-O, et al. Efficient and secure outsourcing of genomic data storage. BMC Med Genomics. 2017;10: 46.

7. Lauter K, López-Alt A, Naehrig M. Private Computation on Encrypted Genomic Data. Progress in Cryptology - LATINCRYPT 2014. Springer International Publishing; 2015. pp. 3–27.

8. Kubiritova Z, Gyuraszova M, Nagyova E, Hyblova M, Harsanyova M, Budis J, et al. On the critical evaluation and confirmation of germline sequence variants identified using massively parallel sequencing. J Biotechnol. 2019;298: 64–75.

9. Ayday E, Raisaro JL, Hengartner U, Molyneaux A, Hubaux J-P. Privacy-Preserving Processing of Raw Genomic Data. In: Garcia-Alfaro J, Lioudakis G, Cuppens-Boulahia N, Foley S, Fitzgerald WM, editors. Data Privacy Management and Autonomous Spontaneous Security. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. pp. 133–147.

10. Minarik G, Repiska G, Hyblova M, Nagyova E, Soltys K, Budis J, et al. Utilization of benchtop next generation sequencing platforms ion torrent PGM and MiSeq in noninvasive prenatal testing for chromosome 21 trisomy and testing of impact of in silico and physical size selection on its analytical performance. PLoS One. 2015;10: e0144811.

11. Pös O, Budiš J, Szemes T. Recent trends in prenatal genetic screening and testing. F1000Res. 2019;8. doi:10.12688/f1000research.16837.1

12. Pös O, Budis J, Kubiritova Z, Kucharik M, Duris F, Radvanszky J, et al. Identification of Structural Variation from NGS-Based Non-Invasive Prenatal Testing. Int J Mol Sci. 2019;20. doi:10.3390/ijms20184403

13. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443.

14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a

tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

15. Ayday E, Raisaro JL, Hubaux J-P, Rougemont J. Protecting and evaluating genomic privacy in medical tests and personalized medicine. Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society. ACM; 2013. pp. 95–106.

16. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15: 409–421.

17. Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Deriving genomic diagnoses without revealing patient genomes. Science. 2017;357: 692–695.