Privacy in Biomedical Data Science 3/24

Objectives of the Lecture

- Explain security in the field of bioeconomy
- Identify cyber-biosecurity vulnerabilities
- Gain an understanding of privacy in biomedicine and bioeconomy

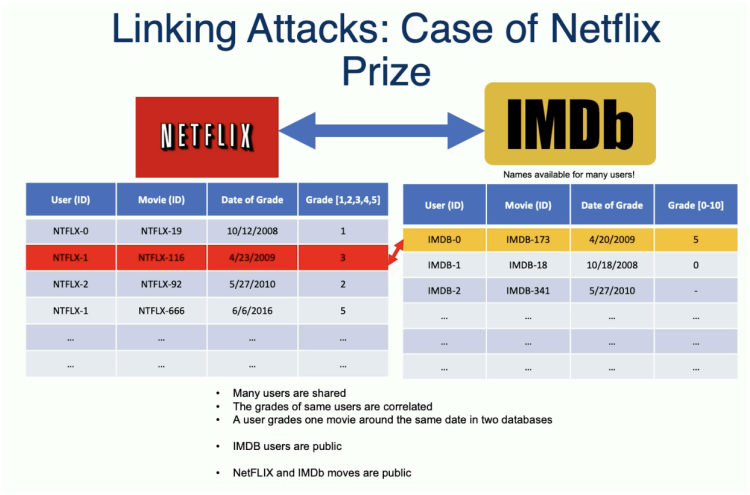
Key Concepts and Definitions

- Privacy-preserving File Formats (pBAM), stands for Privacy-preserving Binary Alignment Mapping
- Sanitization is removing a variant from BAM files, to prevent breaches in privacy and security, and block unauthorized access
- Bioeconomy: economic activity involving the use of biotechnology and biomass in the production of goods, services, currently worth around 30 trillion
- Bioconvergence: multidisciplinary approach in the life sciences that combines the disciplines of biotechnology, engineering, and computing
- Cyber-Biosecurity: assessing vulnerabilities, developing counter-strategies, promoting policies that defend biological systems, data, and technologies from cyberattacks
- Digital Twins: a virtual representation of an object/system (example: test bird strikes on virtual engines instead of real ones)
- Privacy: confidentiality, secrets, right to be left alone, anonymity, the right to decide how the world perceives you
- Invasion of privacy: someone taking private information and disclosing it publicly
- Open source intelligence (OSINT)
- Big Data: 5 Vs (volume, velocity, variety, veracity, value)
- The EU Data Act is a regulatory framework that enhances user control over data, ensures fair data sharing, promotes interoperability, and prevents monopolization by large companies while allowing government access in emergencies.

Main Content/Topics

Linking Attacks

Netflix linking attacks exploited viewing patterns and metadata to de-anonymize users by correlating leaked or publicly available data, such as IMDB ratings and reviews, with pseudonymized Netflix records.

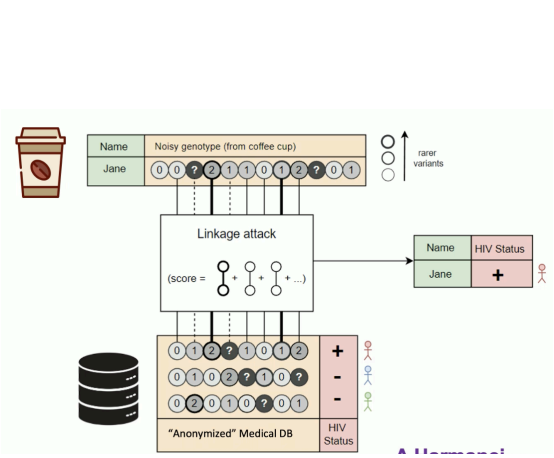


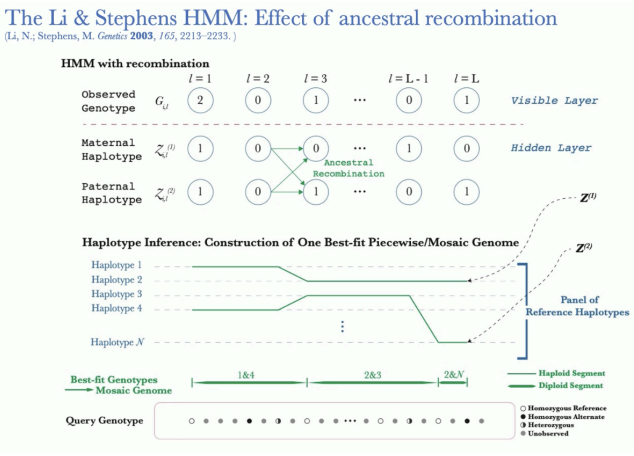
Researchers demonstrated that attackers could re-identify users based on timestamps, viewing histories, or behavioral patterns, posing serious privacy risks by exposing personal habits and preferences. **These vulnerabilities underscored the importance of proper data sanitization in cybersecurity, as inadequate anonymization left users susceptible to re-identification despite privacy safeguards.**

Example of linking attacks in genomic/medical context

Similarly, genomic linkage attacks exploit patterns in noisy SNP (Single Nucleotide Polymorphism) datasets to re-identify individuals and infer sensitive genetic traits. Even when datasets are anonymized or contain errors, attackers can cross-reference them with public genomic databases to link individuals to their genetic profiles. Decreasing SNP information is correlated with an increasing number of equally likely hits. This raises serious privacy concerns, as it enables the exposure of health risks, ancestry, and other personal attributes without consent.

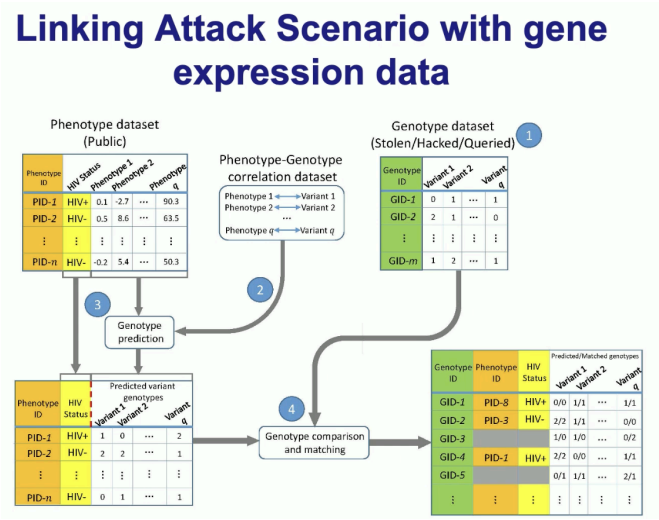
A specific example given in class was using the noisy SNPs that could be isolated from DNA on a coffee cup to identify someone's HIV status:





The Li & Stephen's HMM is a model that can be used to describe paths through genomic databases. The Viterbi algorithm can be used to find the optimal path through the genomic database. As such, the HMM is a useful tool for studying linking attacks, as it can model the and identify patterns similar between the databases.

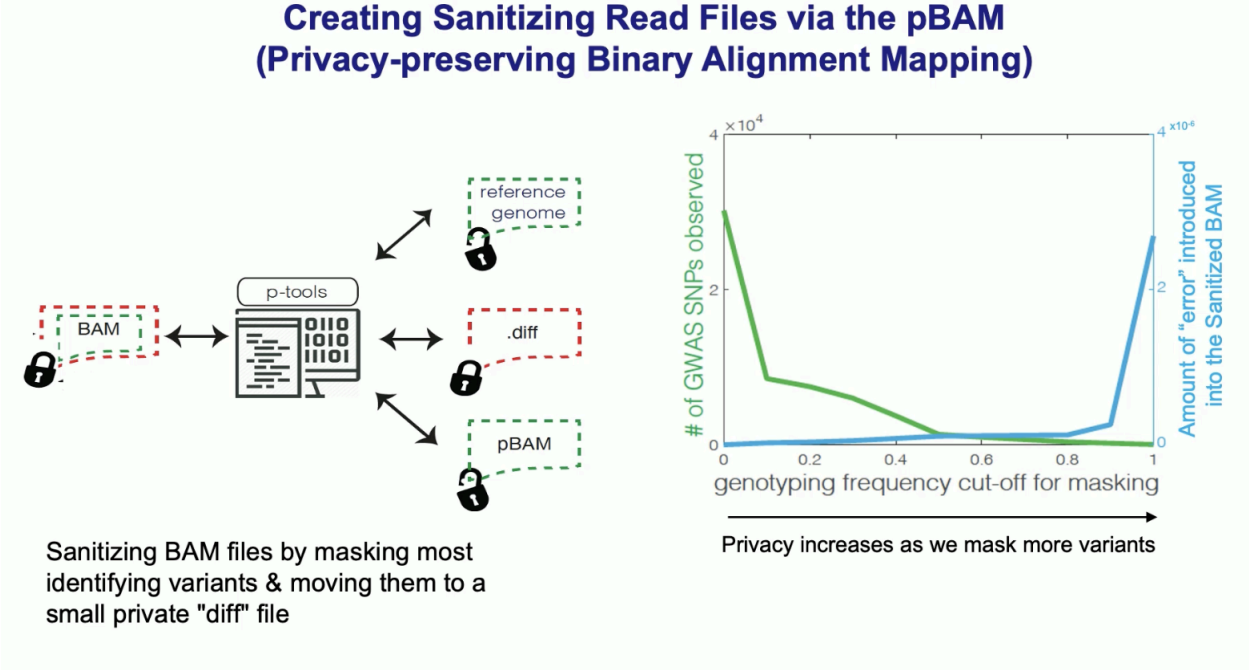
Another example given was linking individuals/HIV status to gene expression data:



Sanitizing BAM files

A way to create privacy-preserving file types (i.e. pBAM). Sanitization would remove sensitive or private information before sharing or distributing them. This sanitization process ensures that personal, medical, or otherwise confidential data (like sample identifiers or metadata) are

stripped from the file, while still retaining the essential alignment information for analysis.



Discussion/Comments

- I wish we covered more guidelines like HIPAA and GDPR
- I would have liked a bigger emphasis on technical solutions (e.g., federated learning, edge computing)
- Sanitization of genomic files is important when sharing data. It is often needed in order to comply with privacy regulations, such as HIPAA in the United States or GDPR in Europe.
- The bankruptcy of 23andMe has intensified debates over the ownership and privacy of genetic data. Over 15 million customers' DNA information are potentially at stake during asset liquidation. The transfer of genetic information to new (and dangerous) entities could alter existing privacy agreements. This situation underscores the open question of who owns their genetic data, and what privacy means in the biomedical world.

List all suggested reading here and please answer:

Gürsoy, G., Li, T., Liu, S. et al. Functional genomics data: privacy risk assessment and technological mitigation. Nat Rev Genet 23, 245–258 (2022). https://doi.org/10.1038/s41576-021-00428-7 (Up to the *Cryptographic approaches* heading)

Are the readings for the class useful? If so, are the specific subsections useful or would change. If not, are there other references you could suggest? Please suggest one.

Yes, especially the paper that Dr. Gerstein suggested at the end of the lecture. In this paper, the sections “Secure sharing of functional genomics data” and subsection “Redacted BAMs” were particularly relevant

Other Suggest references for many of the key concepts

Malakar, Y., Lacey, J., Twine, N.A. et al. Balancing the safeguarding of privacy and data sharing: perceptions of genomic professionals on patient genomic data ownership in Australia. Eur J Hum Genet 32, 506–512 (2024). https://doi.org/10.1038/s41431-022-01273-w

Ellen Wright Clayton, Barbara J Evans, James W Hazel, Mark A Rothstein, The law of genetic privacy: applications, implications, and limitations, Journal of Law and the Biosciences, Volume 6, Issue 1, October 2019, Pages 1–36, https://doi.org/10.1093/jlb/lsz007

Abraham P Schwab, Hung S Luu, Jason Wang, Jason Y Park, Genomic Privacy, Clinical Chemistry, Volume 64, Issue 12, 1 December 2018, Pages 1696–1703, https://doi.org/10.1373/clinchem.2018.289512

Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J. P., Malin, B. A., & Wang, X. (2015). Privacy in the Genomic Era. ACM computing surveys, 48(1), 6. https://doi.org/10.1145/2767007

Li and Stephens HMM:
Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, 165(4), 2213–2233. https://doi.org/10.1093/genetics/165.4.2213