Privacy-Preserving Storage, Sharing, and Analysis for Genomics Data

We are in an era of next-generation sequencing data for genomics. Sharing this vast trove of data is essential for advancing biomedical research while posing significant privacy challenges. The first critical barrier is achieving practical solutions for data ownership and integrity. To this end, we developed a private blockchain network to store genomic variants and reference-aligned reads on-chain. We also established various file formats to reduce leakage during sharing. Our SAMchain approach stored large-scale genomics by minimizing the data inserted into the blockchain using reference-based compression and indexing techniques. Next, we developed algorithms to carry out privacy-preserving transformations to functional genomics data, sanitizing the private variants. Specifically, we created a privacy-preserving file format for raw sequence alignment maps (called pBAM). Finally, we investigated the reads that map to the microbiome from raw human functional genomics data. We used various machine-learning approaches to infer private information about individuals from these microbial mappings. All of our work is added to the pre-existing open-source software associated with SAM, BAM, and CRAM tools.

Appendix A - Reference

[1] Joly, Y, Dyke SOM, Knoppers, BM and Pastinen, T. 2016. Are data sharing and privacy protection mutually exclusive? Cell 167:1150–1154. PMID: 27863233

[2] Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! Genome Biology 12:125. PMCID: PMC3245608

[3] Rodriguez LL, Brooks LD, Greenberg JH, Green ED. 2013. The Complexities of Genomic Identifiability. Science 339:275–276. PMID: 23329035

[4] Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. Nat Rev Genet 15:409–421. PMCID: PMC4151119

[5] Canela-Xandri O, Rawlik K, Tenesa A. 2018. An atlas of genetic associations in UK Biobank. Nat Genet. 50(11):1593-1599. PMID: 30349118

[6] GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45(6):580-5. PMCID: PMC4010069

[7] The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74. PMCID: PMC4750478

[8] The Cancer Genome Atlas (TCGA). The Cancer Genome Atlas Research Network. National Institutes of Health; http://cancergenome.nih.gov/

[9] https://www.washingtonpost.com/news/morning-mix/wp/2018/10/17/the-culprits-name-remains-unknown-but-he-licked-a-stamp-and-now-his-dna-stands-indicted/?utm_term=.25eba675732b

[10] Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 112(22):E2930-8. PMCID: PMC4460507

[11] Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science 339:957–959. PMCID: PMC4423787

[12] Sotiriou C, Pusztai L. 2009. Gene-expression signatures in breast cancer. New England Journal of Medicine 360:790–800. PMID: 19228622

[13] The ENCODE Project Consortium. 2012. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489:57–74. PMCID: PMC3439153

[14] Harmanci A, Gerstein M. 2016. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nature Methods. 13(3):251-6. PMCID: PMC4834871

[15] Harmanci A, Gerstein M. 2018. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun. 9(1):2453. PMCID: PMC6015012

[16] Schadt EE, Woo S, Hao K. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. Nat Genet. 44(5):603-8. PMID: 22484626
[17] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. Science. 339(6117):321-4. PMID: 23329047
[18] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. 4(8):e1000167. PMCID: PMC2516199
[19] Im HK, Gamazon ER, Nicolae DL, Cox NJ. 2012. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet. 90(4):591-8. PMCID: PMC3322234
[20] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20:1297-303. PMCID: PMC2928508
[21] Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics. 43:11.10.1-11.10.33. PMCID: PMC4243306
[22] Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research. 21(6):974-84. PMCID: PMC3106330
[23] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84. PMCID: PMC4197822
[24] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333-i339. PMCID: PMC3436805
[25] Harewood L, Kishore K, Eldridge MD, Wingett S, Pearson D, Schoenfelder S, Collins VP, Fraser P. 2017. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangementsand copy number variation in human tumours. Genome Biology. 18(1):125. PMCID: PMC5488307
[26] Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ozadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F. 2018. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. (10):1388-1398. PMCID: PMC6301019
[27] Chakraborty A, Ay F. 2017. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. [Epub ahead of print]. PMID: 29048467
[28] Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics. 38(11):1341-7. PMID: 17033624
[29] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak M, Gaffney D, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology. 17:13-14. PMCID: PMC4728800
[30] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biology. 9(9):R137. PMCID: PMC2592715
[31] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 9(10):999-1003. PMCID: PMC3816492

[32] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 3(1):95-8. PMCID: PMC5846465
[33] Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research. 24(6):999-1011. PMCID: PMC4032863
[34] Yan KK, Lou S, Gerstein M. 2017. MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS Comput Biol. 13(7):e1005647. PMCID: PMC5546724
[35] Yan KK, Yardimci GG, Yan C, Noble WS, Gerstein M. 2017. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics. 33(14):2199-2201. PMCID: PMC5870694
[36] Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, Thistlethwaite W, Subramanian SL, Milosavljevic A, Gerstein M. exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst. 2019 Apr 24;8(4):352-357.e3. PMID: 30956140
[37] Liang C, Li Y, Luo J. 2016. A Novel Method to Detect Functional microRNA Regulatory Modules by Bicliques Merging. IEEE/ACM Trans Comput Biol Bioinform. 13(3):549-556. PMID: 27295638