# Experience on Privacy Preserving File Formats focused on Functional Genomics and Gene Network Representations

# Previous experience in functional genomics privacy-preserving file formats

Functional genomics experiments have a somewhat more nuanced position with respect to privacy than traditional DNA sequencing. Functional genomics experiments are often performed to understand the biology behind phenotypes, such as different diseases, as opposed to DNA sequencing which is performed solely for genotyping. In fact, many of the conclusions we infer from these experiments are not tied to the identity of individuals but represent universal statements about biology (e.g., genes up-regulated in cancer). However, by virtue of the experimental procedure, the raw sequences contain the genetics variants of the research participants. On the other hand, the knowledge of these genetic variants are often not necessary to calculate key quantities from the functional genomics data, such as gene expression values or transcription factors binding peaks.

Privacy concerns around functional genomics data due to its dependence on next-generation sequencing are typically addressed by increasingly restrictive access provisions limited to a few authorized parties. These restrictions are in tension with the need for large-scale multi-modal analyses that integrate public and private data. In particular, large-scale consortia such as IGVF that aim to generate and disseminate data to a wider community outside the membership of the consortium itself is in immediate need of different modes of raw data sharing beyond controlled-access. We have previously developed an alternate approach: pTools, a theoretical framework and open-source implementation for privacy-preserving transformations that allow standard functional genomics analyses to be performed openly on private data without compromising security. We have shown that the utility loss from privacy-preserving transformation is less than the differences between two biological replicates of a functional genomics sample. We also developed pTools for the use of 10x single-cell RNA-Seq data and showed that downstream analysis such as read counts and cell type clustering are not affected by this transformation1.

pTools takes a reference aligned functional genomics BAM file and creates a (a) privacy-preserving BAM and a (b) .diff file, which contains the differences between BAM and pBAM.

**pBAM details:** The privacy of pBAM is provided by masking the variant information from BAM attributes such as CIGAR, SEQ, alignment score, MD:Z tag and distance to reference. This is done by converting the CIGAR strings to perfectly mapped values, transforming to SEQ to the sequence of that location in the reference genome, converting the alignment score to the perfectly mapped alignment scores, removing the variants from MD:Z tags and converting all the distance to reference (NM tag in BAM files) into 0. This guarantees that variant callers or manual human inspection cannot tell the difference between a pBAM read and the sequence of the reference genome of the read location. pTools is a flexible software that can generate pBAMs regardless of the used reference genome or the aligner. It provides support for single-cell RNA-Seq, ChIP-Seq, RNA-Seq and ATAC-Seq. We also plan on expanding its support to other functional genomics datasets that will be produced in the consortium. pBAM file format is based on the existing file format system SAM/BAM/CRAM. This will allow users to use the existing functional genomics pipelines with pBAMs as input, without exposing sensitive genotype information. Users can also treat pBAM files as BAMs and use existing tools to parse and analyze pBAMs.

**.diff details:** With the motivation of keeping the size of controlled-access file formats relatively small, we include only the differences between the BAM and pBAM file in the diff file by avoiding printing any sequence information of the reads that can be found in the reference human genome. The diff files are private files that require special permission to access. A user is able to recreate the original BAM file when they have access to the diff file by using pTools  that can convert pBAM + pdiff + reference genome into the original BAM (or SAM/CRAM) file.

# Previous experience in gene network representations

The Gerstein Lab has extensive experience in biological network science and has developed a network analysis platform (http://networks.gersteinlab.org). This platform encompasses tools to determine small-scale network motifs such as feed-forward loops and feedback loops as well as large-scale structures such as overall network hierarchies, center points of networks, bottlenecks of networks, and so forth. These have been published in numerous analyses including identifying enriched network motifs with Loregic2, cross-species network clustering with OrthoClust3, and calculating the impact of conserved or species specific regulatory networks on gene expression with DREISS4. Additional network analysis methods have been formulated in a web-accessible network toolkit called TYNA5. This platform has been used to analyze the human regulatory network, the network associated with cancer, the phosphorylation network in yeast, the yeast regulatory network, and other model organism networks6. We have performed extensive comparisons between these regulatory networks and also compared regulatory networks with networks in other contexts such as governmental hierarchies, assembly line control flows, and the call graph structure of the Linux operating system7. We have published many comparative network papers including a review exploring comparing networks across disciplines8. Such an approach enables the transfer of mathematical formalisms from disparate disciplines to help better describe and understand complex biological phenomena.. The Gerstein lab has previous experience in mining the yeast regulatory network9,10, finding the occurrence of various motifs used by the phosphorylation network and the transcriptional regulatory network. We have also recently published a review on how network analysis is a great unifier in biomedical data science11. The IGVF has proposed a heavy investment in projects that will uncover the gene expression networks around systems undergoing a biological transition.   The precise modeling of these experiments and subsequent analysis will vary quite a bit on what is proposed by the awardees, but we envision a layered structure that links a series of datasets across a transition.

Moreover, constructed networks can help infer the direct and indirect effects of genomic variants. We have made use of network properties such as centrality to evaluate the functional significance of genomic variants, revealing that the correlation between them depends on the network type14. Network hierarchy represents another useful concept for predicting the impacts of genomic variants, with the perturbation of elements at the upper layers (such as master regulators) causing more widespread effects than the perturbation of elements at the lower layers, which are more localized. We have developed different methods for determining network hierarchies, such as the HirNet method, which was applied to compare the hierarchies of the phosphorylation and TF binding networks6. Genomic variants can also lead to disruptions of network connections. As a result, some recurrent patterns may be perturbed, such as TFs that frequently co-regulate target genes. We have developed the DiNeR method for identifying such changes and analyzing their consequences to downstream gene expression programs15. On a larger scale, some network perturbations may propagate to cause major network rewiring. We have developed the TopicNet method to measure such rewiring in transcriptional regulatory networks16. We have also applied this idea to study network rewiring in cancer cells, as part of our efforts in producing a general resource for cancer research based on ENCODE data17.

1 Gursoy et al. Data sanitization to reduce private information from functional genomics. Cell, in press (2020)

2 Wang, D. et al. Loregic: a method to characterize the cooperative logic of regulatory factors. PLoS Comput Biol **11**, e1004132, doi:10.1371/journal.pcbi.1004132 (2015).

3 Yan, K. K. et al. OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol **15**, R100, doi:10.1186/gb-2014-15-8-r100 (2014).

4 Wang, D., He, F., Maslov, S. & Gerstein, M. DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. PLoS Comput Biol **12**, e1005146, doi:10.1371/journal.pcbi.1005146 (2016).

5 Yip, K. Y., Yu, H., Kim, P. M., Schultz, M. & Gerstein, M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics **22**, 2968-2970, doi:10.1093/bioinformatics/btl488 (2006).

6 Cheng, C. et al. An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. Genome Biol **16**, 63, doi:10.1186/s13059-015-0624-2 (2015).

7 Yan, K. K., Fang, G., Bhardwaj, N., Alexander, R. P. & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. Proceedings of the National Academy of Sciences of the United States of America **107**, 9186-9191, doi:10.1073/pnas.0914771107 (2010).

8 Yan, K. K. et al. Cross-Disciplinary Network Comparison: Matchmaking Between Hairballs. Cell Syst **2**, 147-157, doi:10.1016/j.cels.2016.02.014 (2016).

9 Wang, D., Yan, K. K., Rozowsky, J., Pan, E. & Gerstein, M. Temporal Dynamics of Collaborative Networks in Large Scientific Consortia. Trends Genet **32**, 251-253, doi:10.1016/j.tig.2016.02.006 (2016).

10 Yan, K. K. & Gerstein, M. The spread of scientific information: insights from the web usage statistics in PLoS article-level metrics. PloS one **6**, e19917, doi:10.1371/journal.pone.0019917 (2011).

11 Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**, 1344-1349, doi:10.1126/science.1158441 (2008).

12 Borneman, A. R. et al. Divergence of transcription factor binding sites across related yeast species. Science **317**, 815-819, doi:10.1126/science.1140748 (2007).

13 McGillivray, P.  et al. Network analysis as a grand unifier in biomedical data science. Annual Reviews of Biomedical Data Science **1**, 153-180

14 Khurana, E. et al. Interpretation of genomic variants using a unified biological network approach. Plos CB 9, e1002886 (2013)

15 Zhang et al. DiNeR: a Differential graphical model for analysis of co-regulation network rewiring. BMC Bioinformatics 21, 281 (2020)

        16 Lou et al. TopicNet: a framework for measuring transcriptional regulatory network change. Bioinformatics **36**, i474-i481 (2020)

17 Zhang et al. An integrative ENCODE resource for cancer genomics. Nat Coomun 11, 3696 (2020)