***Accomplishments***

We set up a systematic way of interacting with our collaborator in University of California at Berkeley. To this end, we have been having bimonthly ongoing phone calls and 2 in-person meetings. We contacted Global Alliance for Genomic Health Consortium (GA4GH) to build collaborations. We are in communication with ENCODE consortium as well as GA4GH for a standardized use of our developed file formats.

***What were the major goals and objectives of the project?***

The goals of this project include development of theoretical frameworks (Aim 1) and software tools (Aim 2) for quantifying private information leakage from RNA-sequencing datasets.

 ***What was accomplished under these goals?***

***Aim 1: Development of a Statistical Formalism for Leakage from QTL Sets***

We formulated information theory-based measures for estimating sensitive information leakage from gene expression values, signal tracks as well as raw alignment files of RNA-Seq data. Leakage from gene expression using eQTL dataset is computed using PrivaSeq and published [1]. Berkeley is working on a tool that can quantify leakage from gene expression using sQTL dataset. Yale has completed a tool called PrivaSig to calculate the sensitive information leakage from signal tracks of RNA-Seq and this tool has been accepted for publication [2]. We are currently developing tools to estimate the information leakage from raw alignment files of RNA-Seq data.

***Aim 2: Instantiating the Linking Attacks***

Yale developed practical ways of instantiating linking attacks using eQTL dataset, signal profiles and raw alignment reads. Berkeley is working on a linking attack based on sQTL dataset. These enable us to test the extent of risks around the linking attacks on real datasets. We have been communicating with ENCODE consortium for a possible collaboration as they started acquiring tissue samples from individuals. We are establishing connections for running our software tools on ENCODE consortium software pipelines. We will make our tools part of the standardized pipelines that ENCODE consortium distributes. This will increase the visibility of our tools significantly. We are also in communication with GA4GH to standardize our file formats and tools. We attended the 2017 GA4GH plenary meeting as well as the accompanying privacy meeting of GenoPri. We are planning to attend 2018 plenary meeting GenoPri as well.

***Aim 3: Building Privacy Enhancing File Formats***

We have created tools that can convert leaky signal profiles into leak free signal profiles. We also developed privacy enhancing file format for raw alignment files of RNA-Seq data. We believe this file format will be very helpful to the biomedical research community, as currently raw alignment files require special access. We developed tools for experimentalists to convert their leaky alignment files into leak-free, high utility files that can be publicly shared. We believe this new file format system will also enhance the reproducibility and will allow researcher to collaborate in a rapid manner.

 ***What opportunities for training and professional development did the
project provide?***

Former trainee Dr. Arif Harmanci and current trainee Dr. Gamze Gursoy proposed a session to the Pacific Symposium in Biocomputing (PSB) conference on Biological Data Privacy and their proposal is accepted. They will co-chair this session and this will help postdoctoral researcher Dr. Gursoy to be recognized in the biomedical data privacy field and establish herself as a rising academic in this field. Both of them submitted abstracts to ISMB 2018 with this project and Dr. Gursoy is also attending the BD2K special session and networking event at ISMB this year. ***How were the results disseminated to communities of interest?***

The current software tools and results can be accessed through privaseq.gersteinlab.org, privasig.gersteinlab.org and privaseq3.gersteinlab.org.

We are currently working on consolidating all the tools under the name of ptools and create a meta-website for researchers to access. ***What do you plan to do during the next reporting period to accomplish
the goals and objectives?***

We will continue developing theoretical frameworks for protecting datasets. We will generalize the formalism that we developed for SNVs to structural variants and continue working on splicing.

We will also work on instantiating privacy attacks focusing on gene expression extremities and proteomics data.

We will generalize our privacy-enhancing file formats for other type of functional genomics data such as ChIP-Seq and Hi-C.

 ***Products***

N/A ***publications, conference papers, and presentations***

We made a poster presentation in the Pacific Symposium in Biocomputing in Hawaii in January 3-7.

We published the tools PrivaSeq and PrivaSig[ 1,2].

We presented a poster at the Inaguration of the Yale Center for Biomedical Data Science,

Trainee Dr. Gamze Guraoy travelled to Switzerlan for a privacy workshop and also visited cancer researcher Prof. Mark Rubin’s lab at the University of Bern to give a seminar on our file format and tools.

We also presented our project at iDASH 2017.

 ***website(s) or other Internet site(s)***

All the data and software can be accessed through privaseq.gersteinlab.org, privasig.gersteinlab.org and privaseq3.gersteinlab.org ***technologies or techniques***N/A. ***other products, such as data or databases, physical collections, audio
or video products, software, models, educational aids or curricula,
instruments or equipment, research material, interventions (e.g.,
clinical or educational), or new business creation.***

N/A

 ***Participants and Other Collaborating Organizations***

Steven Brenner from University of California at Berkeley ***Impact***

Our project has very high impact. ***Changes in approach and reasons for change***

N/A ***Actual or anticipated problems or delays and actions or plans to resolve them***N/A