A. OVERALL COVER PAGE

Project Title: Methods and Software to Enhance Genomic Privacy and Sharing of RNA-Seq Data	
Grant Number: 5U01EB023686-02	**Project/Grant Period:** 09/23/2016 - 06/30/2019
Reporting Period: 09/23/2016 - 06/30/2017	**Requested Budget Period:** 07/01/2017 - 06/30/2018
Report Term Frequency: Annual	**Date Submitted:**
Program Director/Principal Investigator Information: MARK BENDER GERSTEIN , PHD AB **Phone number:** 203- 432-6105 **Email:** mark.gerstein@yale.edu	**Recipient Organization:** YALE UNIVERSITY YALE UNIVERSITY OFFICE OF SPONSORED PROJECTS PO BOX 208327 NEW HAVEN, CT 065208327 **DUNS:** 043207562 **EIN:** 1060646973A1 **RECIPIENT ID:**
Change of Contact PD/PI: N/A	
Administrative Official: SUSAN HEDLEY Office of Sponsored Projects P.O. Box 208327 New Haven, CT 065208327 **Phone number:** 12037854689 **Email:** OSP@YALE.EDU	**Signing Official:** SUSAN HEDLEY Office of Sponsored Projects P.O. Box 208327 New Haven, CT 065208327 **Phone number:** 12037854689 **Email:** OSP@YALE.EDU
Human Subjects: No	**Vertebrate Animals:** No
hESC: No	**Inventions/Patents:** No

B. OVERALL ACCOMPLISHMENTS

B.1 WHAT ARE THE MAJOR GOALS 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.

The milestones for these goals are listed below:
AIM1: * Year 1 – Development of mathematical formalism for quantifying information leakage & privacy attacks and anonymization of the data.
We will also develop anonymization strategies for protecting data at different risk levels that the users deem acceptable for their scenario.

AIM1 and AIM2: * Years 1 & 2 – Development of practical software instantiating the above formalisms.
Software development will be aimed at making the formalisms developed in year 1 practical for everyone's use. We will distribute the software from github where it will be open to public for download and comments.

AIM1 and AIM2: * Year 2 - Large-Scale Deployment of the Risk Management and Anonymization Formalism
We will deploy the implemented software on projects like TCGA, ENCODE, GTex, GSP.

AIM2: * Years 2 & 3 - Develop file formats systematically removing private information (Integration of the Anonymization Formalism with Existing File Formats)
We will develop new data file formats that enable efficient distribution of the anonymized datasets. These data formats will be supported by our software packages for easy accession and processing.

AIM2: * Year 3 - Instantiate other sources of extremity into practical software
We have shown that extremity is a simple yet very effective concept that can be used to breach privacy in linking attacks. We will study different extension of extremity in different data types and scenarios and implement these attacks into the software package.

B.1.a Have the major goals changed since the initial competing award or previous report?

No

B.2 WHAT WAS ACCOMPLISHED UNDER THESE GOALS?

File uploaded: accomplishments.pdf

B.3 COMPETITIVE REVISIONS/ADMINISTRATIVE SUPPLEMENTS

For this reporting period, is there one or more Revision/Supplement associated with this award for which reporting is required?

No

B.4 WHAT OPPORTUNITIES FOR TRAINING AND PROFESSIONAL DEVELOPMENT HAS THE PROJECT PROVIDED?

File uploaded: Individual Developement Plan.pdf

B.5 HOW HAVE THE RESULTS BEEN DISSEMINATED TO COMMUNITIES OF INTEREST?

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

B.6 WHAT DO YOU PLAN TO DO DURING THE NEXT REPORTING PERIOD TO ACCOMPLISH THE GOALS?

We will continue developing theoretical frameworks for protecting datasets. In particular, in the next year, we plan to continue work on all aims of the grant, focusing more on Aim 2, developing practical software and file formats and ways of protecting privacy. We will also generalize the formalism that we developed for SNVs to structural variants and splicing.

We will also generalize the first aim the grant, which focuses on SNVs, to a general system for quantifying leaks and for developing practical attacks, now focusing on SVs and other types of variants such as splicing variants.

At the current reporting period, we have started setting up the project and the databases we will use. We have started extending the information theoretic formalisms to new data types. We are listing major activities

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

We have started formulating information theoretic measures for estimating sensitive information leakage from QTL datasets. We have published PrivaSeq tool for computing sensitive information leakage from gene expression datasets[1]. Our privacy formalism was covered in the original Aim 1 of the grant. We, luckily, were able to get this published soon after it was announced that the grant would be funded. Our major finding is that gene expression datasets can be used to characterize sensitive information in linking attacks.

Aim 2: Development of Software Tools for Instantiating and Simulating Attacks

We have been working on developing practical ways of instantiating linking attacks. These have enabled us to test the extent of risks around the linking attacks on real datasets. We have started simulating the attacks on RNA-seq datasets. We have been communicating with GA4GH consortium for a possible collaboration. We are establishing connections for running our software tools on GA4GH consortium software pipelines. We will make our tools part of the standardized pipelines that GA4GH consortium distributes. This will increase the visibility of our tools significantly. We also published another paper where we outline basic rules for building supplementary materials [2]. A big way formats and summary analysis are presented is through the supplements of papers and the structuring of them thereof; we present some proposals for structuring supplements. We have also been discussing implementing these with various journal editors (eg for Genome Biology and Nature).

REFERENCES

1. Harmanci, A. & Gerstein, M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat. Methods* **13,** 251–256 (2016).

2. Greenbaum, D., Rozowsky, J., Stodden, V. & Gerstein, M. Structuring supplemental materials in support of reproducibility. *Genome Biol.* **18,** 64 (2017).

Training and professional development

All graduate students supported by an NIH award at Yale University are required to create an individual development plan. Students provide updates on their IDP activities as part of their annual thesis committee meetings, and documentation is retained by the students' graduate programs.

All postdoctoral trainees at Yale University are required to create an individual development plan and to provide annual progress reports for review by, and discussion with, the faculty mentor. Progress reports are then submitted to the Yale Office for Postdoctoral Affairs as a condition of the trainee's reappointment by this Office.

C. OVERALL PRODUCTS

C.1 PUBLICATIONS

Are there publications or manuscripts accepted for publication in a journal or other publication (e.g., book, one-time publication, monograph) during the reporting period resulting directly from this award?

Yes

Publications Reported for this Reporting Period

Public Access Compliance	Citation
Complete	Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nature methods. 2016 March;13(3):251-6. PubMed PMID: 26828419; PubMed Central PMCID: PMC4834871.

C.2 WEBSITE(S) OR OTHER INTERNET SITE(S)

Category	Explanation
Data or Databases	All the data and software can be accessed through privaseq.gersteinlab.org

C.3 TECHNOLOGIES OR TECHNIQUES

NOTHING TO REPORT

C.4 INVENTIONS, PATENT APPLICATIONS, AND/OR LICENSES

Have inventions, patent applications and/or licenses resulted from the award during the reporting period?

No

C.5 OTHER PRODUCTS AND RESOURCE SHARING

NOTHING TO REPORT

D. OVERALL PARTICIPANTS

D.1 WHAT INDIVIDUALS HAVE WORKED ON THE PROJECT?

Commons ID	S/K	Name	Degree(s)	Role	Cal	Aca	Sum	Foreign Org	Country	SS
MGERSTEIN	Y	Gerstein, Mark Bender	AB,PHD	PD/PI	0	1	0			NA
AHARMANCI	N	Harmanci, Arif	MS,BS,PHD	Postdoctoral Scholar, Fellow, or Other Postdoctoral Position	1	0	0			NA

Glossary of acronyms:
S/K - Senior/Key
DOB - Date of Birth
Cal - Person Months (Calendar)
Aca - Person Months (Academic)
Sum - Person Months (Summer)
Foreign Org - Foreign Organization Affiliation
SS - Supplement Support
RE - Reentry Supplement
DI - Diversity Supplement
OT - Other
NA - Not Applicable

D.2 PERSONNEL UPDATES

D.2.a Level of Effort

Will there be, in the next budget period, either (1) a reduction of 25% or more in the level of effort from what was approved by the agency for the PD/PI(s) or other senior/key personnel designated in the Notice of Award, or (2) a reduction in the level of effort below the minimum amount of effort required by the Notice of Award?

No

D.2.b New Senior/Key Personnel

Are there, or will there be, new senior/key personnel?

No

D.2.c Changes in Other Support

Has there been a change in the active other support of senior/key personnel since the last reporting period?

D.2.d New Other Significant Contributors

Are there, or will there be, new other significant contributors?

No

D.2.e Multi-PI (MPI) Leadership Plan

Will there be a change in the MPI Leadership Plan for the next budget period?

NA

E. OVERALL IMPACT

E.1 WHAT IS THE IMPACT ON THE DEVELOPMENT OF HUMAN RESOURCES?

Not Applicable

E.2 WHAT IS THE IMPACT ON PHYSICAL, INSTITUTIONAL, OR INFORMATION RESOURCES THAT FORM INFRASTRUCTURE?

Our project has very high impact on considerations of data dissemination and safe release of personalized genomic data.

E.3 WHAT IS THE IMPACT ON TECHNOLOGY TRANSFER?

Not Applicable

E.4 WHAT DOLLAR AMOUNT OF THE AWARD'S BUDGET IS BEING SPENT IN FOREIGN COUNTRY(IES)?

NOTHING TO REPORT

F. OVERALL CHANGES

F.1 CHANGES IN APPROACH AND REASONS FOR CHANGE

Not Applicable

F.2 ACTUAL OR ANTICIPATED CHALLENGES OR DELAYS AND ACTIONS OR PLANS TO RESOLVE THEM

NOTHING TO REPORT

F.3 SIGNIFICANT CHANGES TO HUMAN SUBJECTS, VERTEBRATE ANIMALS, BIOHAZARDS, AND/OR SELECT AGENTS

F.3.a Human Subjects

No Change

F.3.b Vertebrate Animals

No Change

F.3.c Biohazards

No Change

F.3.d Select Agents

No Change

G. OVERALL SPECIAL REPORTING REQUIREMENTS

G.1 SPECIAL NOTICE OF AWARD TERMS AND FUNDING OPPORTUNITIES ANNOUNCEMENT REPORTING REQUIREMENTS

NOTHING TO REPORT

G.2 RESPONSIBLE CONDUCT OF RESEARCH

Not Applicable

G.3 MENTOR'S REPORT OR SPONSOR COMMENTS

Not Applicable

G.4 HUMAN SUBJECTS

G.4.a Does the project involve human subjects?

No

G.4.b Inclusion Enrollment Data

Not Applicable

G.4.c ClinicalTrials.gov

Does this project include one or more applicable clinical trials that must be registered in ClinicalTrials.gov under FDAAA?

G.5 HUMAN SUBJECTS EDUCATION REQUIREMENT

Are there personnel on this project who are newly involved in the design or conduct of human subjects research?

G.6 HUMAN EMBRYONIC STEM CELLS (HESCS)

Does this project involve human embryonic stem cells (only hESC lines listed as approved in the NIH Registry may be used in NIH funded research)?

No

G.7 VERTEBRATE ANIMALS

Does this project involve vertebrate animals?

No

G.8 PROJECT/PERFORMANCE SITES

Organization Name:	DUNS	Congressional District	Address
Primary: Yale University	043207562	CT-003	Yale University 266 Whitney Avenue, 432A New Haven CT 065208114
University of California - Berkeley	124726725	CA-013	461A Koshland Hall Berkeley CA 947203102

G.9 FOREIGN COMPONENT

No foreign component

G.10 ESTIMATED UNOBLIGATED BALANCE

G.10.a Is it anticipated that an estimated unobligated balance (including prior year carryover) will be greater than 25% of the current year's total approved budget?

Yes

Estimated unobligated balance: 0

G.10.b Provide an explanation for unobligated balance:
We anticipate a large carry forward in the budget. This is mainly related to administrative issues, i.e., the first year is funded for nine months, as opposed to 12, even though we got full budget. Also, it took quite a while to set the grant up at Yale and subsequently to get our sub-contract set up at Berkeley. Scientifically, we feel that carrying the money forward make sense, since we are initially focusing on the formalism, but we anticipate needing more work with software development for the latter stages of the grants involving practical file formats, particularly after we get some feedback from our initial interaction with consortia such as GA4GH and various journals or repositories.

G.10.c If authorized to carryover the balance, provide a general description of how it is anticipated that the funds will be spent
The carryover balance will be used in the following years mainly for development of databases and software tools. We are expecting some of the cost also to be spent on cloud resources that we will use to test our software packages. We are also foreseeing that some funding may be necessary for accessing some of the proprietary databases.

G.11 PROGRAM INCOME

Is program income anticipated during the next budget period?

No

G.12 F&A COSTS

Not Applicable

Budget Justification

Mark Gerstein, Ph.D. PI (.3 summer months). Dr. Gerstein is the Albert Williams Professor of Biomedical Informatics. His lab (http://gersteinlab.org) was one of the first to perform integrated data mining on functional genomics data and to do genome-wide surveys. His tools for analyzing motions and packing are widely used. Most recently, he has designed and developed a wide array of databases and computational tools to mine genome data in humans, as well as in many other organisms. He has worked extensively in the 1000 genomes project in the SV and FIG groups. He also worked in the ENCODE pilot project and currently works extensively in the ENCODE and modENCODE production projects. He is also a co-PI in DOE KBase and the leader of the Data Analysis Center for the NIH exRNA consortium. In these roles Dr. Gerstein has designed and developed a wide array of databases and computational tools to mine genomic data in humans as well as in many other organisms. He will lead the overall informatics effort in the project.

Dr. Arif Harmanci, Ph.D., Assoc. Research Scientist (12 calendar months). Dr. Harmanci has extensive experience with bioinformatic approaches to genome-wide analysis and a strong background in scientific computation. As part of his PhD thesis, he developed advanced methods for RNA secondary structure prediction. In the Gerstein laboratory, he has developed new algorithms to identify transcription factor binding peaks from ChIP-Seq data. He is currently working on transcriptome, epigenome, and variant analysis of several large scale RNA-seq, DNA-seq, and ChIP-Seq datasets that include the Geuvadis dataset (RNA-Seq on 500 individuals), TCGA dataset, and ENCODE datasets. He will work on the analysis proposed in the grant under the direction of Dr Gerstein. He will work on developing the information theoretic quantification of sensitive individual characterizing information.

Timur R. Galeev, Ph. D., Postdoctoral Associate (7 calendar months). Dr. Galeev has a strong expertise in scientific computation. Before joining the Gerstein lab at Yale University, he obtained his Ph.D. degree (2014) from Utah State University. His Ph.D. research was in the field of theoretical and computational physical chemistry, focused both on applications of modern electronic structure methods and development of new theoretical tools and models of molecular structure and bonding. He is currently working on analysis of functional genomics data. Dr. Galeev will work on developing new file formats that enable efficient and effective distribution of molecular phenotyping datasets in a privacy-aware manner.

Fringe benefits are calculated at the rate of 32.3% for the PI and the Associate Research Scientist and Postdoctoral Associates according to the University guidelines.

TRAVEL

As this is a collaborative project, we are budgeting considerable funds for travel between sites. Here we are requesting incremental funds for each of the FTEs for airfare, lodging and meal expenses to attend scientific meetings annually that benefit the project. In particular, the travel will include at least 1 trip per year to a scientific meeting of in genomics and bioinformatics such as the ISMB or CSHL Biology of Genomes.

SUPPLIES

We are budgeting an incremental amount of supplies for the individuals named above. This supplies budget will be used to cover computer supplies for them, covering such expenses as: diskettes, tapes, and other miscellaneous computer parts (e.g. replacing worn out surge suppressors), software upgrades, web hosting and "cloud computing fees, and reprint charges. These items are needed to complete the proposed research and will solely benefit this project.

SUBCONTRACT

University of California, Berkeley – see separate budget and justification.

INDIRECT COST

Indirect costs are calculated at Yale’s federally negotiated rate of 67.5% of modified total direct costs. DHHS agreement dated 02/16/2017.