AWARD NUMBER: W81XWH-18-1-0722

TITLE: Cell Communication in Antiestrogen Resistance

PRINCIPAL INVESTIGATOR: Robert Clarke, PhD, DSc

CONTRACTING ORGANIZATION: University of Minnesota, Twin Cities, Minneapolis, MN

REPORT DATE: October 2022

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2022	Annual	`15Sep2021-14Sep2022

4. TITLE AND SUBTITLE

Cell Communication in Antiestrogen Resistance

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-18-1-0722

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Robert Clarke & Yue Wang

E-Mail: clarker@umn.edu & yuewang@vt.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Hormel Institute,
Regents of the University of Minnesota
and
Virginia Polytechnic Inst & State University

8. PERFORMING ORGANIZATION REPORT NUMBER

2

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

More women die from the estrogen receptor positive (ER+) breast cancer subtype than from any other. The proportion of early ER+ recurrences (=5 years since diagnosis) approaches that for all triple-negative breast cancers alone. Late recurrences (>5 years after diagnosis), the result of dormancy, are most common in ER+ disease and can arise decades after the initial diagnosis. Since recurrent breast cancers have escaped the effects of endocrine therapies, and are lethal, we will study endocrine resistance (Tamoxifen; Fulvestrant). Our primary objective is to identify what drives breast cancer growth and determine how to stop it. We will learn about why some breast cancers are aggressive and others are indolent, and why/how some breast cancers lay dormant for years and then re-emerge.

15. SUBJECT TERMS

Breast cancer, drug resistance, admixing, ecology, multiscale modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	18	19b. TELEPHONE NUMBER *(include area code)*

TABLE OF CONTENTS

1. Introduction

~70% of newly diagnosed breast cancers are ER+ [1]. Many of these women die because metastatic ER+ disease becomes treatment resistant. Resistance is multiscale, i.e., evident at many levels, with genetic, cellular, and phenotypic features (including intratumor heterogeneity; ITH), all are molecularly manifested, and functionally realized, as networked changes in the transcriptome and proteome. We will take a systems biology approach to portray the proteome and transcriptome topology of treatment-induced adaptive remodeling of cell admixtures in vitro and in vivo. Overarching goals are to understand the principles of this remodeling and uncover the mechanisms that confer endocrine resistance in breast cancer, leading to new treatment strategies.

NB: As mentioned in the prior report we have lost a significant portion of time due to COVID-19 restrictions at University of Minnesota and Virginia Tech. Please also note that the contacting PI Dr. Clarke moved to University of Minnesota at the beginning of October 2020. Thus, we have fallen behind in some of the wet laboratory work due to the move from DC to MN during the current pandemic. We kept some of the *in silico* work moving forward and we submitted and published some new work in this area. The wet laboratory studies are now progressing and we have other manuscripts submitted and/or in preparation.

2. Keywords

Drug resistance, admixing, ecology, multiscale modeling

3. Accomplishments

A) Major goals (and related subtasks) of the project from approved SOW:

SPECIFIC AIM 1 (specified in proposal)	Timeline	Site 1	Site 2	Percent complete	Date completed (if 100%)
Major Task 1 (Aim 1a)	Months				
Subtask 1: Determine prevalence of **R** and/or **P** cells in **S** populations	1-6	Dr. Clarke Dr. Sengupta			
Subtask 2: Determine the effects of different **S**:**R** ratios on response to TAM and ICI in MCF-7, LCC, T47D, and ZR-75-1 **S** and **R** matched cell models *in vitro*	1-18	Dr. Clarke Dr. Sengupta			
Milestone(s) Achieved: Identified effects of **S**:**R** ratio on responsiveness to TAM and ICI *in vitro* in multiple breast cancer cell models and identified optimal admix ratios for *in vivo* studies					
Local IACUC approval (annual renewal required only – approval for studies already in place)	1	Dr. Clarke		Approved (100%)	Nov 2022
Local IRB approval (add this award as an exemption for use of existing data – no new clinical data will be generated in this BT#2)	1	Dr. Clarke		In preparation	
Subtask 3: Determine the effects of different **S**:**R** ratios on response to TAM and ICI in **S**:**R** matched cell models *in vivo* (models and admixes guided by the optimal* results in Major Task 1/Subtask 2) It is difficult to provide direct numbers until the *in vitro* work is completed. A standard design for a single would include the following (n=15/group as in application): R cells alone ± ICI (15+15=30) S cells alone ± ICI (15+15=30) R+S cells at a single ration ± ICI (15+15=30) Total = 90/experiment We may do 3 such experiments over the 18-month period for 270 mice.	6-24	Dr. Clarke Dr. Sengupta		10%	
Milestone(s) Achieved: Identified effects of **S**:**R** ratio on responsiveness to TAM and ICI *in vivo*					
Major Task 2 (Aim 1b)					
Subtask 1: Determine the role of GJIC in the ability of **R** to make **S** cells resistant to TAM and ICI in MCF-7, LCC, T47D and ZR-75-1 matched cell models *in vitro* (guided by the optimal experimental conditions from Aim 1a)	6-18	Dr. Clarke Dr. Sengupta		20%	

Subtask 2: Determine the role of microvesicles and protein secretion (transwell) in the ability of R to make S cells resistant to TAM and ICI in MCF-7, LCC, T47D and ZR-75-1 matched cell models *in vitro* (informed by the optimal experimental conditions identified in Aim 1a)	6-18	Dr. Clarke Dr. Sengupta		10%	
Milestone(s) Achieved: Identified role GJIC, microvesicles and protein secretion (transwell) in the ability of **R** to make **S** cells resistant to drug and how this is affected by different **S**:**R** ratios. Identified conditions to allow design and execution of *in vivo* studies with guggulsterone and/or GW4869 (experiments will be done if supported by data and if time permits)					
Subtask 3: Collect and store materials (*e.g.*, cell lysates) from optimal conditions for omics studies in Aim 2	1-24	Dr. Clarke Dr. Sengupta			
SPECIFIC AIM 2 (specified in proposal)	**Timeline**	**Site 1**	**Site 2**		
Major Task 3 (Aim 2a)					
Subtask 1: Collect RNA and protein from the materials stored from Aim 1a (this will be done as the optimal experiments are identified above)	1-24	Dr. Clarke Dr. Sengupta	Dr. Wang	75% Site 2	
Subtask 2: Perform array and proteome data collection, processing of raw data from Major Task 3/Subtask 1 (above), and initial/in-depth analyses (*e.g.*, CAM, kDDN)	1-24	Dr. Clarke Dr. Sengupta	Dr. Wang	75% Site 2	
Milestone(s) Achieved: Create initial signaling maps of what is communicated by **R** to **S** to confer resistance and how this is affected by different **S**:**R** ratios					
Major Task 4 (Aim 2b)					
Subtask 1: Build initial mathematical models of cell population remodeling dynamics (*in vitro* and *in vivo* data)	4-24	Dr. Bansal		75%	
Subtask 2: Build final mathematical models of cell population remodeling dynamics (*in vitro* and *in vivo* data)	24-36	Dr. Bansal			
Milestone(s) Achieved: Identified how endocrine therapies and the starting ratios of **S**:**R** cells affects population responses to treatment					
Major Task 5 (Aim 2c)					
Subtask 1: Use the data from Aims 1 and 2 to design and execute novel drug combination and scheduling studies *in silico* (mathematical modeling), *e.g.*, ICI+DNMTi	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	30% Site 2	
Subtask 2: Design and execute novel drug combination and scheduling studies *in vitro* using the predictions in Major Task 5/Subtask 1	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	30% Site 2	
Milestone(s) Achieved: Identified novel optimized (activity vs. toxicity) combination regimens *in vitro*.					
Subtask 3: A small number of predictions from the *in vitro* modeling in	18-36	Dr. Clarke Dr. Sengupta	Dr. Wang	30% Site 2	

Major Task 5/Subtask 2 will be tested *in vivo* (we anticipate completing ~5 such animal studies) It is difficult to provide direct numbers until the *in vitro* work is completed. A standard design for a single would include the following (n=15/group as in application): R cells alone + Vehicle (15) S cells alone + Vehicle (15) R cells alone + Drug A and + Drug B (15+15=30) S cells alone + Drug A and + Drug B (15+15=30) R+S cells at a single ratio with Vehicle, + Drug A and + Drug B (15+15+15=45) Total = 135/experiment We may do 4 such experiments over the funding period (n=540 maximum number mice).		Dr. Bansal			
Milestone(s) Achieved: Identified novel optimized (activity vs. toxicity) combination regimens *in vivo*.					
SPECIFIC AIM 3 (specified in proposal)	**Timeline**	**Site 1**	**Site 2**		
Major Task 6 (Aim 3a)					
Subtask 1: Initial CAM and kDDN modeling of microarray data from human tumors (public and in-house datasets); data will be fed back to Aim 2 to increase clinical relevance	1-12		Dr. Wang	100% Site 2	8/31/19
Subtask 2: Update models using outcomes from Aim 2 and study if candidate molecules from Aim 2 are associated with clinical outcome (univariate and multivariate)	12-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	75% Site 2	
Milestone(s) Achieved: Identified clinically relevant molecules associated with ITH and endocrine resistance					
Subtask 4: A small number of predictions from the *in vitro* modeling in Major Task 5/Subtask 3 will be tested *in vivo* (~5 such experiments will be done) A small number of predictions from the in vitro modeling in Major Task 5/Subtask 2 will be tested in vivo (we anticipate completing ~5 such animal studies) It is difficult to provide direct numbers until the in vitro work is completed. A standard design for a single would include the following (n=15/group as in application): R cells alone ± Drug A (15+15=30) S cells alone ± Drug A (15+15=30)	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	50% Site 2	

R cells alone ± Drug B (15+15=30) S cells alone ± Drug B (15+15=30) R+S cells at a single ratio + Drug A + Drug B (15+15=30) Total = 150/experiment We may do 3-5 such experiments over the funding period (n=750 maximum number mice).					
Milestone(s) Achieved: Identified novel therapeutic strategies for ER+ breast cancer to prevent, delay or reverse resistance, and do so within minimized toxicity. These insights could be used to design clinical trials to be done outside this research program.					
Major Task 7 (Aim 3b)					
Subtask 1: Test candidate molecules from the model predictions in Aims 2 and 3a. For example, as described in the narrative section, genes upregulated in resistant cells relative to sensitive cells will be overexpressed (cDNA; regulable and/or constitutive promoters) in sensitive cells and knocked down in resistant (RNAi) if their mRNA or protein is still present in sensitive cells. The gene will be knocked out (CRISPR) in resistant cells if the gene is known to be lost or expression is undetectable in sensitive cells. The reverse experiments will be done where a gene is down regulated or lost in resistant cells relative to its expression/presence in sensitive cells.	12-36 months	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	50% Site 2	
Milestone(s) Achieved: Identified mechanistically relevant molecules associated with ITH and endocrine resistance					

B) What was accomplished under these goals? (Site 1: *in vitro* studies at UMN)

Upon moving to The Hormel Institute, University of Minnesota (Austin, MN), we procured the live cell imaging system, Incucyte SX-5. This new equipment is capable of imaging live cells using five different fluorescence channels, up to three fluorescence channels can be used simultaneously in long term experiments time-lapse experiments. The equipment was installed in June 2021 and we have since optimized its use for collecting more *in vitro* data from the admixing studies. Thus, using our fluorescently tagged cells we have re-optimized the imaging condition in our cell culture systems at our new institution and are poised to restart collecting *in vitro* data to complete the studies as proposed.

Significant delays occurred because of closure of laboratories due to COVID-19 related restrictions and re-establishment of our laboratory (site 1) at The Hormel Institute, University of Minnesota, Austin. Drs. Clarke and Sengupta left Georgetown University, Washington DC (original site 1) and moved their laboratories to the University of Minnesota, Austin MN in mid-2020. At the time, both universities were closed dues to covid restrictions and the physical relocation of laboratory equipment alone was delayed by ~ 6 months (for which we received an initial no-cost extension). However, the University of Minnesota opened at first 50%, then 75% and finally 100% capacity over subsequent months, further delaying reinitiating of the research. We had to restart the laboratory and then hire and train new staff. While we are now at 100% capacity, we have lost over 18 months. The studies at Virginia Tech (site 2) were mostly able to continue using the data we had already generated and with other data in the public domain. For some of this work, we were able to use non-breast cancer datasets to build, test and validate some of our new

tools; this is not unusual in the field. Hence, we are largely on-track for the quantitative analysis and tool development work (site 2) and we are working hard to catch up on the wet laboratory work (site 1).

Major Task 1

Subtask 1: Determine prevalence of **R** and/or **P** cells in **S** populations

Subtask 2: Determine the effects of different **S**:**R** ratios on response to TAM and ICI in MCF-7, LCC, T47D, and ZR-75-1 **S** and **R** matched cell models *in vitro*

Subtask 3: Determine the effects of different **S**:**R** ratios on response to TAM and ICI in **S**:**R** matched cell models *in vivo* (models and admixes guided by the optimal* results in Major Task 1/Subtask 2)

It is difficult to provide direct numbers until the *in vitro* work is completed. A standard design for a single would include the following (n=15/group as in application):

R cells alone ± ICI (15+15=30)

S cells alone ± ICI (15+15=30)

R+S cells at a single ration ± ICI (15+15=30)

Total = 90/experiment

We originally planned to do 3 such experiments over the 18-month period for 270 mice. Given the delays, it is likely that some of these will be completed after funding has ended but we are committed to this work and intend to complete as much as possible. We have not started these experiments but have recently received approval from the UMN IACUC (not yet from DoD).

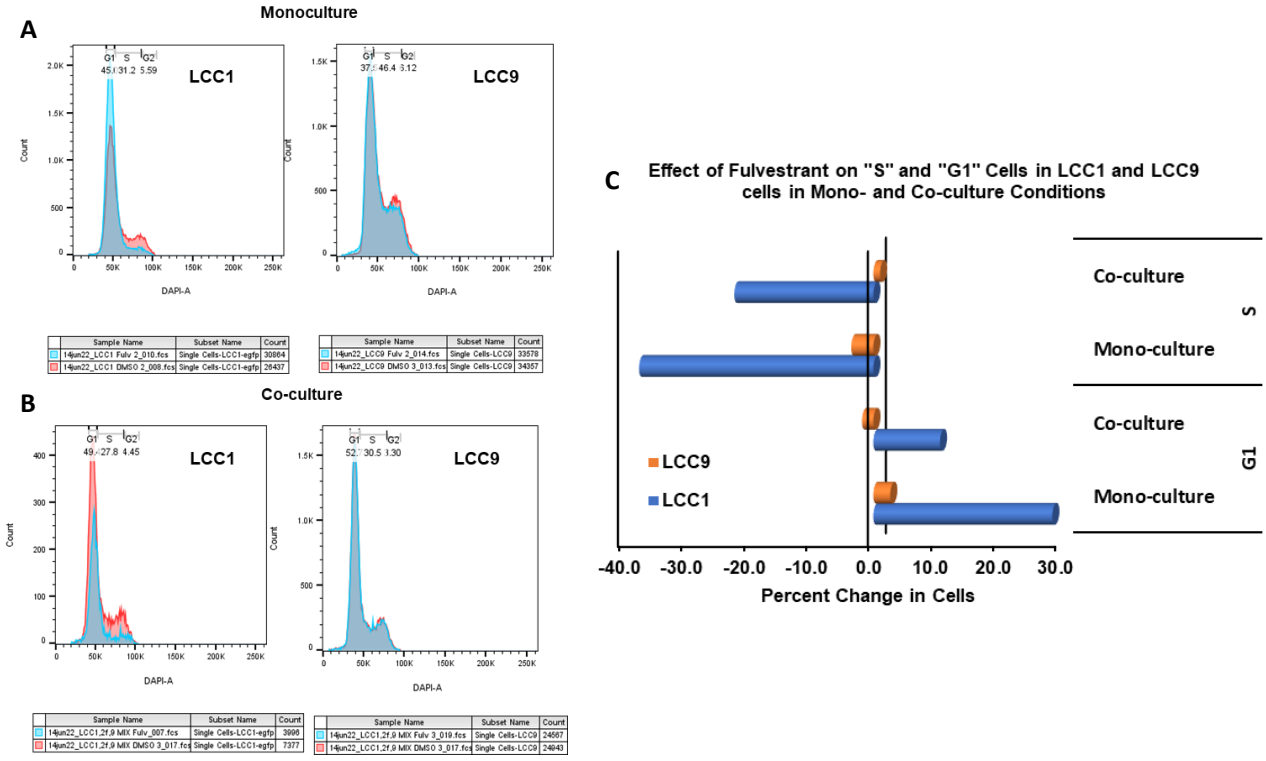


Figure 1. Effect of Fulvestrant treatment on cell cycle in monoculture or co-culture conditions. The LCC1-egfp and LCC9-mcherry cells either monocultured (**A**) or co-culture (**B**) (1:1, ratio) and treated with vehicle or Fulvestrant 500nM for 7 days followed by cell cycle analysis. S= synthesis phase cells, G1= cells in G1 phase. Pink shade=vehicle, blue shade= Fulvestrant 500nM. (**C**) The bar graph shows the percent change in 'S' and 'G1' phase cells in LCC1 and LCC9 after Fulvestrant treatment under monoculture and co-culture conditions.

Progress: We have accomplished subtask 2 (50%) where we have determined effect of different S:R ratios on response to TAM and ICI (please see previous reports).

Upon the re-establishment of the laboratory at The Hormel Institute, University of Minnesota, Austin, we have optimized the growth conditions of the cells. **Figure 1 A and B** shows the cells in different phases of cell cycle after seven (07) days of Fulvestrant (500nM) treatment either under monoculture (**A**)

conditions or co-culture (B) conditions. Quantification of percent change in S phase and G1 cells (**Figure 1C**) following Fulvestrant treatment shows that there is ~37% reduction in S phase LCC1 cells following Fulvestrant treatment in monoculture conditions, whereas ~22% reduction in S phase cells was observed under co-culture conditions. The G1 cells were 30% higher after Fulvestrant treatment in LCC1 under monoculture conditions but there was only 10% increase in G1 LCC1 cells with Fulvestrant treatment when co-cultured with LCC9 cells (**Figure 1C**). No significant changes were noted in LCC9 cells with Fulvestrant treatment under mono- and co-culture conditions. This indicates a diminished inhibitory effect of Fulvestrant on LCC1 cells when the cells are co-cultured with LCC9 cells.

Major Task 2

Subtask 1: Determine the role of GJIC in the ability of **R** to make **S** cells resistant to TAM and ICI in MCF-7, LCC, T47D and ZR-75-1 matched cell models *in vitro* (guided by the optimal experimental conditions from Aim 1a)

Subtask 2: Determine the role of microvesicles and protein secretion

(transwell) in the ability of R to make S cells resistant to TAM and ICI in MCF-7, LCC, T47D and ZR-75-1 matched cell models *in vitro* (informed by the optimal experimental conditions identified in Aim 1a)

Subtask 3: Collect and store materials (*e.g.,* cell lysates) from optimal conditions for omics studies in Aim 2

Progress: We are in the process of collecting the cell lysates and other samples for omics studies.

Major Task 3

Subtask 1: Collect RNA and protein from the materials stored from Aim 1a (this will be done as the optimal experiments are identified above)

Subtask 2: Perform array and proteome data collection, processing of raw data from Major Task 3/Subtask 1 (above), and initial/in-depth analyses (*e.g.,* CAM, kDDN)

Progress: We are in the process of collecting the samples (RNA and protein).

Major Task 4

Subtask 1: Build initial mathematical models of cell population remodeling dynamics (*in vitro* and *in vivo* data)

Subtask 2: Build final mathematical models of cell population remodeling dynamics (*in vitro* and *in vivo* data)

Progress: The initial mathematical models of cell population remodeling dynamics (subtask 1) has been performed (please see previous reports) and a manuscript is currently under review.

Major Task 5

Subtask 1: Use the data from Aims 1 and 2 to design and execute novel drug combination and scheduling studies *in silico* (mathematical modeling), *e.g.,* ICI+DNMTi

Subtask 2: Design and execute novel drug combination and scheduling studies *in vitro* using the predictions in Major Task 5/Subtask 1

Subtask 3: A small number of predictions from the *in vitro* modeling in Major Task 5/Subtask 2 will be tested *in vivo* (we anticipate completing ~5 such animal studies)

Progress: This task will be performed once we have the data from Aim 1 and 2.

Major Task 6

Subtask 1: Initial CAM and kDDN modeling of microarray data from human tumors (public and in-house datasets); data will be fed back to Aim 2 to increase clinical relevance.

Subtask 2: Update models using outcomes from Aim 2 and study if candidate molecules from Aim 2 are associated with clinical outcome (univariate and multivariate)

Subtask 3: A small number of predictions from the *in vitro* modeling in Major Task 5/Subtask 3 will be tested *in vivo* (~5 such experiments will be done)

A small number of predictions from the in vitro modeling in Major Task 5/Subtask 2 will be tested in vivo (we anticipate completing ~5 such animal studies)

It is difficult to provide direct numbers until the *in vitro* work is completed. A standard design for a single would include the following (n=15/group as in application):

R cells alone ± Drug A (15+15=30)
S cells alone ± Drug A (15+15=30)
R cells alone ± Drug B (15+15=30)
S cells alone ± Drug B (15+15=30)
R+S cells at a single ratio + Drug A + Drug B (15+15=30)
Total = 150/experiment
We may do 3-5 such experiments over the funding period (n=750 maximum number mice).

Progress: (Subtask 1 and 2; Site 2: *in silico* studies at VT)

B-1. Between-group normalization of biologically diverse samples

Data normalization is essential to ensure accurate inference and comparability of gene expressions across samples or conditions. Ideally, gene expressions should be rescaled based on consistently expressed reference genes. However, for normalizing biologically diverse samples, most commonly used reference genes have exhibited striking expression variability, and distribution-based approaches can be problematic when differentially expressed genes are significantly asymmetric.

We introduce a Cosine score based iterative normalization (Cosbin) strategy to normalize biologically diverse samples. The between-sample normalization is based on iteratively identified consistently expressed genes, where differentially expressed genes are sequentially eliminated according to scale-invariant Cosine scores. We demonstrate the performance of Cosbin on realistic simulation data sets, followed by the case study on normalizing real biologically diverse samples. Implemented in open-source R scripts and applicable to grouped or individual samples, the Cosbin tool will allow biologists to detect subtle yet important molecular signals across known or novel phenotypic groups.

The R Scripts of Cosbin is freely available at https://github.com/MinjieShen/Cosbin

B-2. swCAM unsupervised sample-wise deconvolution of subtype-specific expressions in individual samples.

Complex biological tissues are often a heterogeneous mixture of several molecularly distinct cell subtypes. Both subtype compositions and subtype-specific expressions can vary across biological conditions. Computational deconvolution aims to dissect patterns of bulk tissue data into subtype compositions and subtype-specific expressions. Existing deconvolution methods can only estimate averaged subtype-specific expressions in a population, while many downstream analyses such as inferring co-expression networks in particular subtypes require subtype expression estimates in individual samples. However, individual-level deconvolution is a mathematically underdetermined problem because there are more variables than observations.

We report a sample-wise Convex Analysis of Mixtures (swCAM) method that can estimate subtype proportions and subtype-specific expressions in individual samples from bulk tissue transcriptomes (**Figure 2**). We extend our previous CAM framework to include a new term accounting for between-sample variations and formulate swCAM as a nuclear-norm regularized matrix factorization problem. We determine hyperparameter values using cross-validation with random entry exclusion and obtain a swCAM solution using an efficient alternating direction method of multipliers. Experimental results on realistic simulation data show that swCAM can accurately estimate subtype-specific expressions in individual samples and successfully extract co-expression networks in particular subtypes that are otherwise unobtainable using bulk data. In two real-world applications, swCAM analysis of bulk RNASeq ROSMAP data from brain tissue of cases and controls with bipolar disorder or Alzheimer's disease identified significant changes in cell proportion, expression pattern and co-expression module in patient neurons. Comparative evaluation of swCAM versus peer methods is also provided. We should clarify that we used ROSMAP data as a testbed to validate the performance of swCAM tool because ROSMAP provides both bulk data, cell-sorted data, and ICH cell proportions, serving as the ground truth.

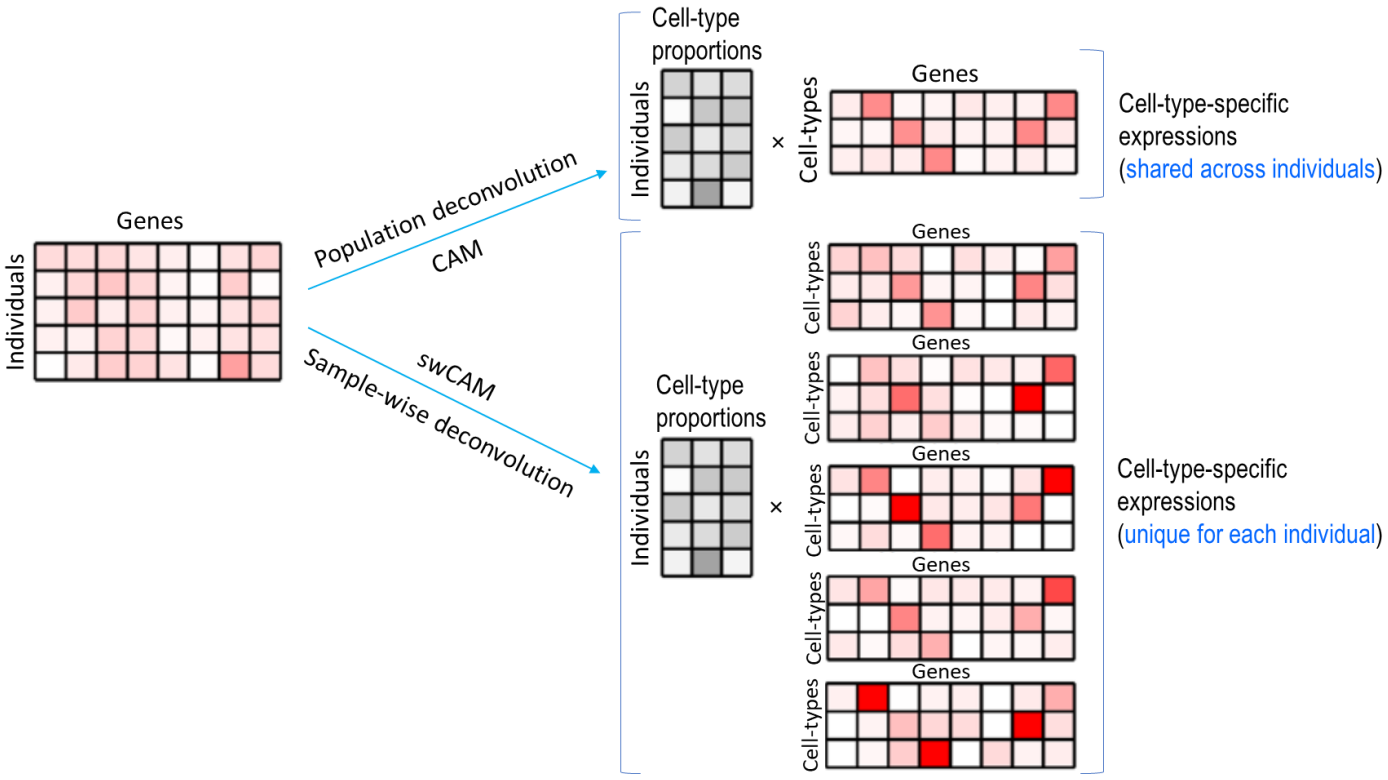


Figure 2. swCAM versus earlier CAM decomposition approach. Given bulk gene expression data from a heterogeneous tissue, earlier CAM approach aims at estimating a matrix of the cell-type proportions of the individuals and a matrix of the cell-type-specific transcriptome in the sample (shared across individuals). In contrast, swCAM aims at estimating a matrix of the cell-type proportions of the individuals and - for each individual—a matrix of the unique cell-type-specific transcriptome of the individuals.

The R Scripts of swCAM is freely available at https://github.com/Lululuella/swCAM

B-3. Comparative assessment and novel strategy on methods for imputing proteomics data

Missing values are a major issue in quantitative proteomics or other molecular expression analysis. While many methods have been developed for imputing missing values in high-throughput proteomics data, a comparative assessment of imputation accuracy remains inconclusive, mainly because mechanisms contributing to true missing values are complex and existing evaluation methodologies are imperfect. Moreover, few studies have provided an outlook of future methodological development. We first re-evaluate the performance of eight representative methods targeting three typical missing mechanisms (**Figure 3**). These methods are compared on both simulated and masked missing values embedded within real proteomics datasets, and performance is evaluated using three quantitative measures (**Figure 4**). We then introduce fused regularization matrix factorization, a low-rank global matrix factorization framework, capable of integrating local similarity derived from additional data types. We also explore a biologically-inspired latent variable modeling strategy - convex analysis of mixtures - for missing value imputation and present preliminary experimental results. While some winners emerged from our comparative assessment, the evaluation is intrinsically imperfect because performance is evaluated indirectly on artificial missing or masked values not authentic missing values. Nevertheless, we show that our fused regularization matrix factorization provides a novel incorporation of external and local information, and the exploratory implementation of convex analysis of mixtures presents a biologically plausible new approach (**Figure 5**).

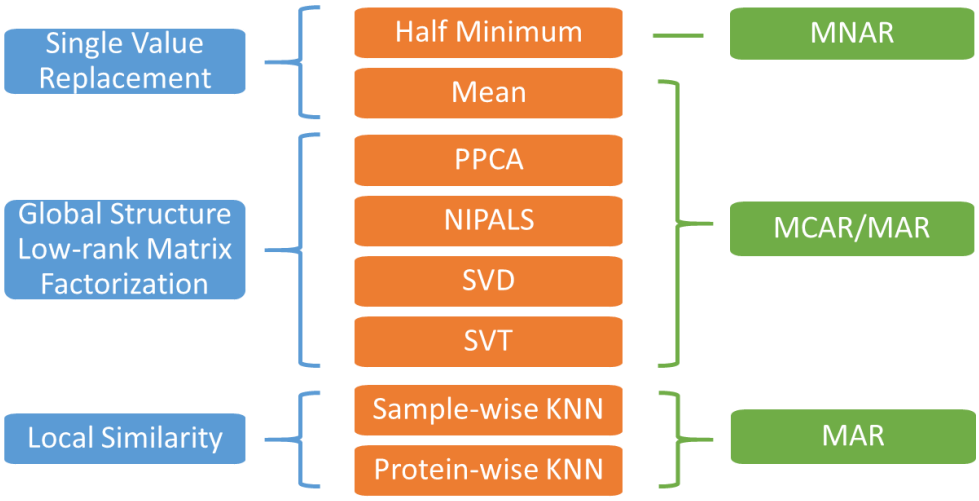


Figure 3. Comparative assessment of eight representative missing value imputation methods, divided into three categories.

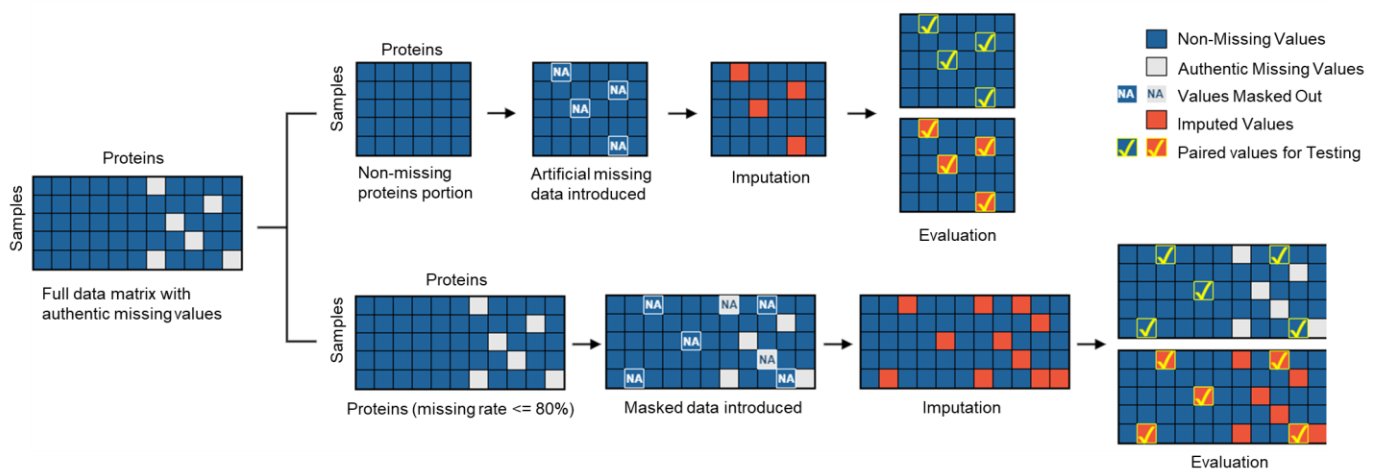


Figure 4. Two-phased workflow of realistic simulation-based assessment on missing value imputation methods.

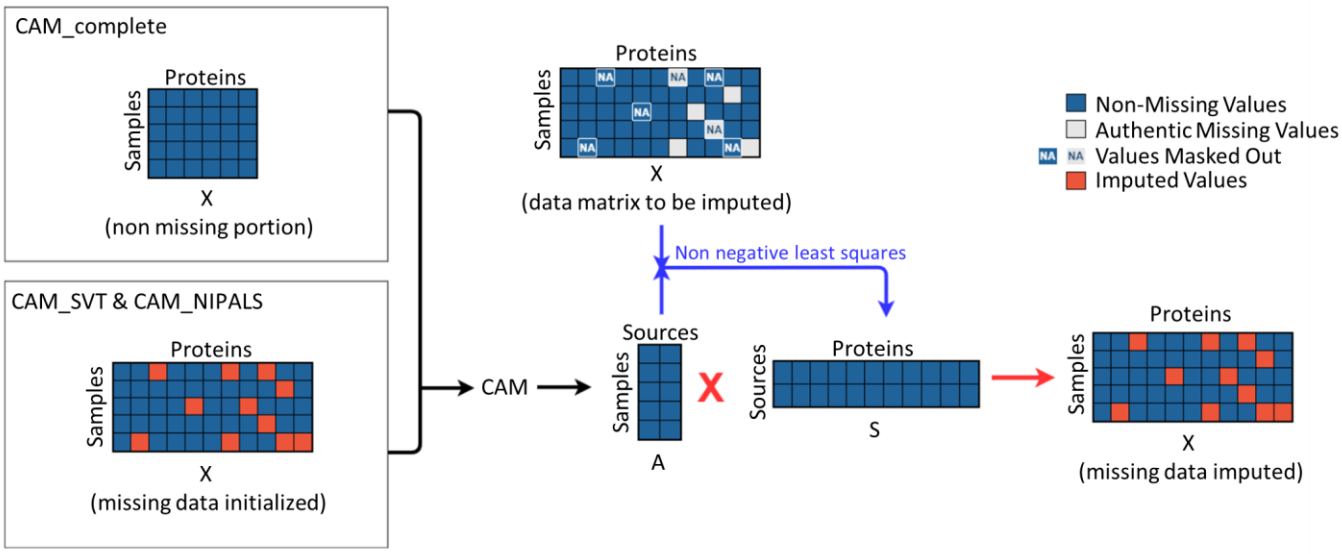


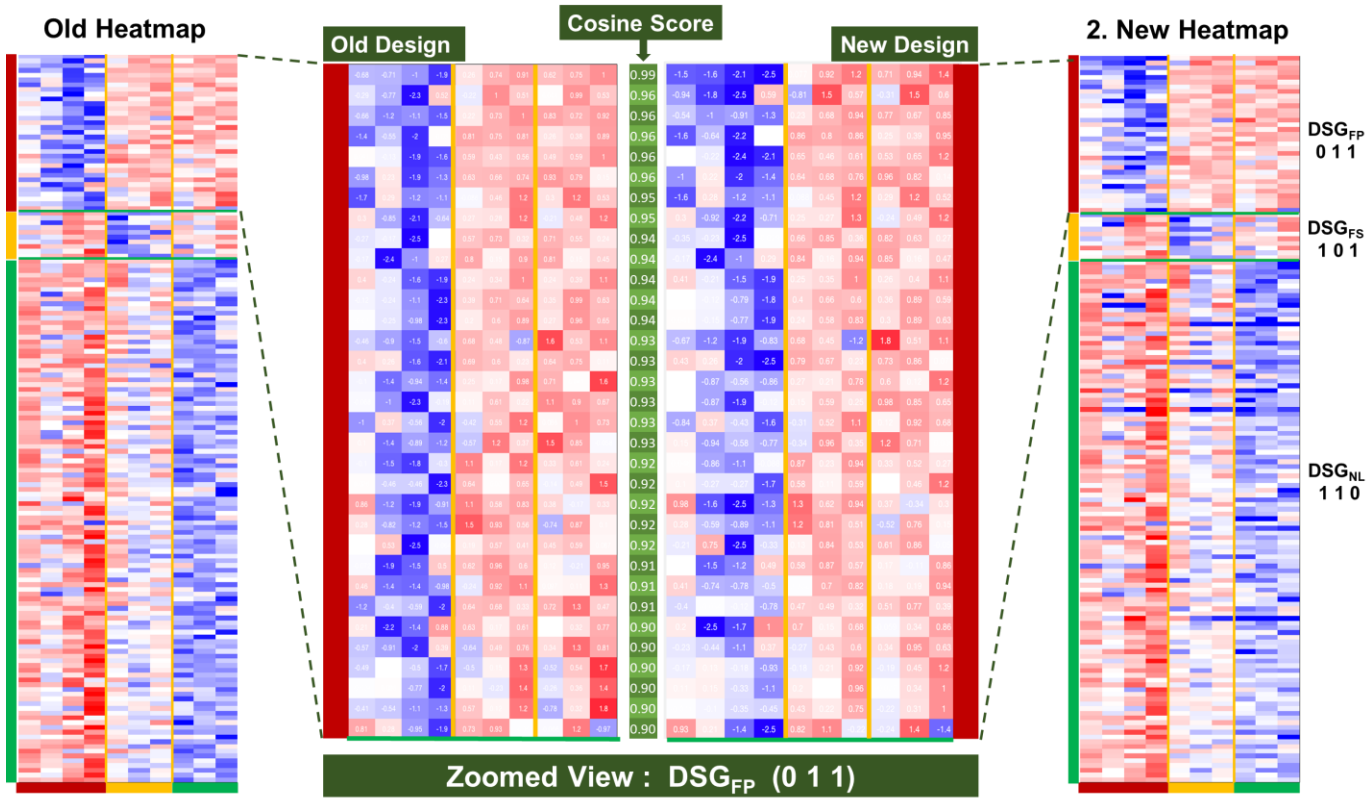
Figure 5. Workflow of the CAM based imputation method with two variant algorithms.

The R script of ProImput freely available at https://github.com/MinjieSh/ProImput

B-4. COTIN: an integrated data-analysis tool suite for study of biologically heterogeneous samples

Data analysis tools are essential to ensure accurate inference of gene expression levels and changes across different conditions. One fundamental analytic task is to detect and visualize phenotypic signature genes, often supported by between-sample normalization and missing value imputation. However, for studying biologically diverse samples, most commonly used data-analysis tools may be problematic when multiple combinatorial biological subtypes are compared, different missing mechanisms are unevenly involved across subtypes, or the amount of asymmetry in differential expression is significant.

We report an integrated data-analysis tool suite – Cosine-score based test, imputation and normalization (COTIN) – specifically designed for studying biologically diverse samples. Based on the Cosine scores of cross-subtype expression patterns, the COTIN tools detect subtype-specific or subtype-combinatorial signature genes using enumerated one-sample tests, impute subtype-specific missing values by proportionally integrating different mechanisms, and normalize gene expressions by eliminating asymmetric differential expressions and identifying consistently expressed genes (**Figure 6**). Implemented in open-source R scripts, COTIN tool suite complements rather than replaces the existing tools and will allow biologists to detect interpretable molecular signals among diverse phenotypic samples more accurately.



.

Figure 6. New design of heatmap showing subtype-combinatorial signature genes (scSGs) across multiple phenotypic subtypes, where the quality of scSGs are reflected with sidebar of cosine scores.

<u>Major Task 7</u>

Subtask 1: Test candidate molecules from the model predictions in Aims 2 and 3a. For example, as described in the narrative section, genes upregulated in resistant cells relative to sensitive cells will be overexpressed (cDNA; regulable and/or constitutive promoters) in sensitive cells and knocked down in resistant (RNAi) if their mRNA or protein is still present in sensitive cells. The gene will be knocked out (CRISPR) in resistant cells if the gene is known to be lost or expression is undetectable in sensitive cells. The reverse experiments will be done where a gene is down regulated or lost in resistant cells relative to its expression/presence in sensitive cells.

Progress: This task will be performed once we have completed Aims 2 and 3a.

C) Opportunities for training and professional development

Two doctoral graduate research assistants are supported in part by this project. In return, they have also contributed to developing data analytics tools and pipelines.

D) How the results were disseminated to communities of interest

Coasbin: The R Scripts of Cosbin is freely available at https://github.com/MinjieShen/Cosbin

swCAM: The R Scripts of swCAM is freely available at https://github.com/Lululuella/swCAM

ProImput: The R script of ProImput freely available at https://github.com/MinjieSh/ProImput.

E) Plans for the final reporting period to accomplish the study goals

We will apply Cosbin, swbCAM, ProImput, and DDN tools to analyze both the public data and the data generated in-house at University of Minnesota (Georgetown University Medical Center) site.

We will jointly interpret the results and develop sensible hypotheses.

We will revise the analytics experimental designs toward more focused testing on the hypotheses.

We will further improve the data analytics tools by incorporating the feedback from UMN team.

4. Impact

What was the impact on the development of the principle discipline(s) of the project?

Interpreting an expression profile of complex tissues requires knowledge of both the relative abundance of the different cell or tissue subtypes and their individual expression patterns. Understanding the relative contribution of individual cell or tissue subtypes in individual samples may illuminate pathophysiologic mechanisms, biologic responses to various stimuli, or transitions in tissue phenotype - especially when the cell-cell and cell-matrix interactions in a complex system are necessary conditions for appropriate cell or tissue function. As a fully unsupervised method, swCAM complements not replaces the existing supervised methods when the required supervising information is less accurate or incomplete, particularly when encountering novel subtypes (e.g., cancer subclones due to mutations) or missing values (*e.g.,* scRNA-seq/snRNA-seq reference has limited genes that can be detected). swCAM has a theoretical advantage in cancer studies, since tumors may have novel cell subtypes that have no reference to be used (e.g. due to novel mutations). As a fully unsupervised method, swCAM is ideally applicable and suitable for deconvolving tumor samples.

Though reference data from single-cell sequencing or sorted cells are available for many, there are some situations where reference data are not available or impossible to generate. For example, new types of tumor cells generated with cancer mutations will be unlikely to have a reference. Single-cell sequencing can only cover those genes with relatively high abundance. Sorted cell data is only available for known cell types that can be sorted. Therefore, low abundance subtypes or genes will not have a proper reference. Cell types that cannot be sorted also will not have an appropriate reference. Additionally, reference-based methods rely heavily on the pre-defined marker genes, which sometimes can be challenging. For example, in early developmental stage, progenitor cells and neuronal cells are highly correlated and different from matured cells. Therefore, classic marker genes may not work well in defining transitional cell types references.

The COTIN tool suite integrates and extends the existing functions in COT, ProImput and Cosbin, to include new functions that tailor COT to detect subtype-combinatorial signature genes, permit ProImput to impute subtype-specific missing values, and allow Cosbin to normalize diverse samples containing missing values. Specifically, we propose combinatorial binary strings to enumerate subtype-combinatorial signature references, new heatmap design to display subtype-combinatorial differential expressions, and integrated imputation to address uneven missing rates and mechanisms across different subtypes. The COTIN tool suite will allow biologists to more accurately detect true molecular signals from biologically diverse samples.

What was the impact on other disciplines?

Nothing to Report.

What was the impact on technology transfer?

Nothing to Report.

What was the impact on society beyond science and technology?

Nothing to Report.

5. Changes/Problems

Nothing to report for the science (no technical or other problems).

Changes in approach and reasons for change

N/A

Actual or anticipated problems or delays and actions or plans to resolve them

N/A

Changes that had a significant impact on expenditures

N/A

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

N/A

Significant changes in use or care of human subjects

Significant changes in use or care of vertebrate animals

Significant changes in use of biohazards and/or select agents

6. Products:

Publications.

[1] Y Wang, DM Herrington, “Machine intelligence enabled radiomics,” ***Nature Machine Intelligence*** (2021) 3:838-839.

[2] Lulu Chen, Yingzhou Lu, Chiung-Ting Wu, Robert Clarke, Guoqiang Yu, Jennifer E Van Eyk, David M Herrington, Yue Wang, “Data-driven detection of subtype-specific differentially expressed genes,” ***Scientific Reports***, vol. 11, 332 (2021).

[3] Ming Fan, Zhenyu Fu, Maosheng Xu, Shiwei Wang, Sangma Xie, Xin Gao, Yue Wang, Lihua Li, “A Deep Matrix Completion Method for Imputing Missing Histological Data in Breast Cancer by Integrating DCE-MRI Radiomics,” ***Medical Physics***, pp. 1-13, 2021.

[4] Lulu Chen, Chiung-Ting Wu, Chia-Hsiang Lin, Rujia Dai, Chunyu Liu, Robert Clarke, Guoqiang Yu, Jennifer E. Van Eyk, David M. Herrington, and Yue Wang, “swCAM: estimation of subtype-specific expressions in individual samples with unsupervised sample-wise deconvolution,” ***Bioinformatics***, R1 submitted, 2021.

[5] Minjie Shen, Yi-Tan Chang, Chiung-Ting Wu, Sarah J. Parker, Georgia Saylor, Yizhi Wang, Guoqiang Yu, Jennifer E. Van Eyk, Robert Clarke, David M. Herrington, and Yue Wang, “Comparative assessment and novel strategy on methods for imputing proteomics data,” ***Scientific Reports***, 12:1067, 2022 | https://doi.org/10.1038/s41598-022-04938-0

[6] Clarke R, Kraikivski P, Jones BC, Sevigny CM, Sengupta S, Wang Y., “A systems biology approach to discovering pathway signaling dysregulation in metastasis” ***Cancer Metastasis Rev****.* 2020 Sep;39(3):903-918. PMID: 32776157

[7] Jones BC, Pohlmann PR, Clarke R, Sengupta S., “Treatment against glucose-dependent cancers through metabolic PFKFB3 targeting of glycolytic flux.” ***Cancer Metastasis Rev***. 2022, Jun; 41(2): 447-458 doi: 10.1007/s10555-022-10027-5.

[8] Wu, C.-T., M. Shen, D. Du, Z. Cheng, S. J. Parker, Y. Lu, J. E. V. Eyk, G. Yu, R. Clarke, D. M. Herrington and Y. Wang (2022). "Cosbin: Cosine score based iterative normalization of biologically diverse samples." ***Bioinformatics Adv*** **3**: vbac076.

Books or other non-periodical, one-time publications.
N/A

Other publications, conference papers, and presentations
N/A

Website(s) or other Internet site(s)
N/A

Technologies or techniques
N/A

Inventions, patent applications, and/or licenses
N/A

Other Products
N/A

7. Participants & Other Collaborating Organizations
The person months reported below are for the timeframe for this technical report, 09/15/2021-09/14/2022.

Name: Robert Clarke
Project Role: Principal Investigator (Initiating PI)
Research Identifier (ORCID ID): 0000-0002-9278-0854
Nearest person months worked: 1.8
Contribution to project: Oversaw the wet lab research, data interpretation and edited manuscripts and project report.

Name: Surojeet Sengupta
Project Role: Co-Investigator
Nearest person months worked: 4.2
Contribution to project: Designed experiments, performed experiments, interpreted data, prepared manuscript and supervised day-to-day experiments performed by a research scientist.

Name: Igor Entin
Project Role: Research Scientist
Nearest person months worked: 3.0
Contribution to project: Performed experiments, analyzed data.

Name: Yue Wang
Project Role: Principal Investigator (Partnering PI)
Research Identifier (ORCID ID): https://orcid.org/0000-0002-1788-1102
Nearest person months worked: 1.0
Contribution to project: Designed data analytics methodology

Name: Chiung-Ting Wu
Project Role: Graduate Research Assistant
Nearest person months worked: 6.0
Contribution to project: Developed data deconvolution tool swCAM

Name: Minjie Shen
Project Role: Graduate Research Assistant
Nearest person months worked: 6.0
Contribution to project: Developed R tools Cosbin and ProImput

Has there been a change in the active other support of the PI or senior personnel since the last reporting period?
Nothing to report.

What other organizations were involved as partners?
Nothing to report (aside from Partnering PI institution, Virginia Tech).

8. Special Reporting Requirements
The Partnering-PI (Yue Wang, PhD) has submitted an independent annual report for this period.

9. Appendices