

AWARD NUMBER: W81XWH-18-1-0722

TITLE: Cell Communication in Antiestrogen Resistance

PRINCIPAL INVESTIGATOR: Robert Clarke, PhD, DSc

CONTRACTING ORGANIZATION: University of Minnesota, Minneapolis, MN

REPORT DATE: October 2021

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 is Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

*Form Approved*  
*OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

<b>1. REPORT DATE</b> October 2021		<b>2. REPORT TYPE</b> Annual		<b>3. DATES COVERED</b> 15Sep2020-14Sep2021	
<b>4. TITLE AND SUBTITLE</b> Cell Communication in Antiestrogen Resistance				<b>5a. CONTRACT NUMBER</b>	
				<b>5b. GRANT NUMBER</b> W81XWH-18-1-0723	
				<b>5c. PROGRAM ELEMENT NUMBER</b>	
<b>6. AUTHOR(S)</b> Robert Clarke & Yue Wang  E-Mail: <a href="mailto:clarker@umn.edu">clarker@umn.edu</a> & <a href="mailto:yuewang@vt.edu">yuewang@vt.edu</a>				<b>5d. PROJECT NUMBER</b>	
				<b>5e. TASK NUMBER</b>	
				<b>5f. WORK UNIT NUMBER</b>	
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b>  The Hormel Institute, Regents of the University of Minnesota & Virginia Polytechnic Inst. & State University				<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>  1	
<b>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>  U.S. Army Medical Research and Development Command Fort Detrick, Maryland 21702-5012				<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>	
				<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>	
<b>12. DISTRIBUTION / AVAILABILITY STATEMENT</b>  Approved for Public Release; Distribution Unlimited					
<b>13. SUPPLEMENTARY NOTES</b>					
<b>14. ABSTRACT</b> 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.					
<b>15. SUBJECT TERMS</b> Breast cancer, drug resistance, admixing, ecology, multiscale modeling					
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT</b>	<b>18. NUMBER OF PAGES</b>	<b>19a. NAME OF RESPONSIBLE PERSON</b>
<b>a. REPORT</b>	<b>b. ABSTRACT</b>	<b>c. THIS PAGE</b>			<b>19b. TELEPHONE NUMBER</b> (include area code)
Unclassified	Unclassified	Unclassified	Unclassified	16	USAMRDC

## **TABLE OF CONTENTS**

<b>1. Introduction</b>	<b>4</b>
<b>2. Keywords</b>	<b>4</b>
<b>3. Accomplishments</b>	<b>5</b>
<b>4. Impact</b>	<b>13</b>
<b>5. Changes/Problems</b>	<b>14</b>
<b>6. Products</b>	<b>15</b>
<b>7. Participants &amp; Other Collaborating Organizations</b>	<b>15</b>
<b>8. Special Reporting Requirements</b>	<b>4</b>
<b>9. Appendices</b>	<b>4</b>

## 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 have managed to keep some of the *in silico* work moving forward and we submitted and published some new work in this area.

## 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%)
<b>Major Task 1 (Aim 1a)</b>	Months				
<b>Subtask 1:</b> Determine prevalence of <b>R</b> and/or <b>P</b> cells in <b>S</b> populations	1-6	Dr. Clarke Dr. Sengupta			
<b>Subtask 2:</b> Determine the effects of different <b>S:R</b> ratios on response to TAM and ICI in MCF-7, LCC, T47D, and ZR-75-1 <b>S</b> and <b>R</b> matched cell models <i>in vitro</i>	1-18	Dr. Clarke Dr. Sengupta			
<b>Milestone(s) Achieved:</b> Identified effects of <b>S:R</b> ratio on responsiveness to TAM and ICI <i>in vitro</i> in multiple breast cancer cell models and identified optimal admix ratios for <i>in vivo</i> studies					
Local IACUC approval (annual renewal required only – approval for studies already in place)	1	Dr. Clarke		Under review at UMN	
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	
<b>Subtask 3:</b> Determine the effects of different <b>S:R</b> ratios on response to TAM and ICI in <b>S:R</b> matched cell models <i>in vivo</i> (models and admixes guided by the optimal* results in Major Task 1/Subtask 2) It is difficult to provide direct numbers until the <i>in vitro</i> 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%	
<b>Milestone(s) Achieved:</b> Identified effects of <b>S:R</b> ratio on responsiveness to TAM and ICI <i>in vivo</i>					
<b>Major Task 2 (Aim 1b)</b>					
<b>Subtask 1:</b> Determine the role of GJIC in the ability of <b>R</b> to make <b>S</b> cells resistant to TAM and ICI in MCF-7, LCC, T47D and ZR-75-1 matched cell models <i>in vitro</i> (guided by the optimal experimental conditions from Aim 1a)	6-18	Dr. Clarke Dr. Sengupta		20%	
<b>Subtask 2:</b> Determine the role of microvesicles and protein secretion	6-18	Dr. Clarke Dr. Sengupta			

(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 <i>in vitro</i> (informed by the optimal experimental conditions identified in Aim 1a)					
<b>Milestone(s) Achieved:</b> 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 <i>in vivo</i> studies with guggulsterone and/or GW4869 (experiments will be done if supported by data and if time permits)					
<b>Subtask 3:</b> Collect and store materials (e.g., cell lysates) from optimal conditions for omics studies in Aim 2	1-24	Dr. Clarke Dr. Sengupta			
<b>SPECIFIC AIM 2 (specified in proposal)</b>	<b>Timeline</b>	<b>Site 1</b>	<b>Site 2</b>		
<b>Major Task 3 (Aim 2a)</b>					
<b>Subtask 1:</b> 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	
<b>Subtask 2:</b> 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	
<b>Milestone(s) Achieved:</b> 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					
<b>Major Task 4 (Aim 2b)</b>					
<b>Subtask 1:</b> Build initial mathematical models of cell population remodeling dynamics ( <i>in vitro</i> and <i>in vivo</i> data)	4-24	Dr. Bansal		75%	
<b>Subtask 2:</b> Build final mathematical models of cell population remodeling dynamics ( <i>in vitro</i> and <i>in vivo</i> data)	24-36	Dr. Bansal			
<b>Milestone(s) Achieved:</b> Identified how endocrine therapies and the starting ratios of S:R cells affects population responses to treatment					
<b>Major Task 5 (Aim 2c)</b>					
<b>Subtask 1:</b> Use the data from Aims 1 and 2 to design and execute novel drug combination and scheduling studies <i>in silico</i> (mathematical modeling), e.g., ICI+DNMTi	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	30% Site 2	
<b>Subtask 2:</b> Design and execute novel drug combination and scheduling studies <i>in vitro</i> using the predictions in Major Task 5/Subtask 1	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	30% Site 2	
<b>Milestone(s) Achieved:</b> Identified novel optimized (activity vs. toxicity) combination regimens <i>in vitro</i> .					
<b>Subtask 3:</b> A small number of predictions from the <i>in vitro</i> modeling in Major Task 5/Subtask 2 will be tested <i>in vivo</i> (we anticipate completing ~5 such animal studies)	18-36	Dr. Clarke Dr. Sengupta Dr. Bansal	Dr. Wang	30% Site 2	

<p>It is difficult to provide direct numbers until the <i>in vitro</i> work is completed. A standard design for a single would include the following (n=15/group as in application):</p> <p>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).</p>					
<p><b>Milestone(s) Achieved:</b> Identified novel optimized (activity vs. toxicity) combination regimens <i>in vivo</i>.</p>					
<p><b>SPECIFIC AIM 3 (specified in proposal)</b></p>	<p><b>Timeline</b></p>	<p><b>Site 1</b></p>	<p><b>Site 2</b></p>		
<p><b>Major Task 6 (Aim 3a)</b></p>					
<p><b>Subtask 1:</b> 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</p>	<p>1-12</p>		<p>Dr. Wang</p>	<p>100% Site 2</p>	<p>8/31/19</p>
<p><b>Subtask 2:</b> Update models using outcomes from Aim 2 and study if candidate molecules from Aim 2 are associated with clinical outcome (univariate and multivariate)</p>	<p>12-36</p>	<p>Dr. Clarke  Dr. Sengupta  Dr. Bansal</p>	<p>Dr. Wang</p>	<p>75% Site 2</p>	
<p><b>Milestone(s) Achieved:</b> Identified clinically relevant molecules associated with ITH and endocrine resistance</p>					
<p><b>Subtask 4:</b> A small number of predictions from the <i>in vitro</i> modeling in Major Task 5/Subtask 3 will be tested <i>in vivo</i> (~5 such experiments will be done)  A small number of predictions from the <i>in vitro</i> modeling in Major Task 5/Subtask 2 will be tested <i>in vivo</i> (we anticipate completing ~5 such animal studies)  It is difficult to provide direct numbers until the <i>in vitro</i> 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)</p>	<p>18-36</p>	<p>Dr. Clarke  Dr. Sengupta  Dr. Bansal</p>	<p>Dr. Wang</p>	<p>50% Site 2</p>	

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).					
<b>Milestone(s) Achieved:</b> 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.					
<b>Major Task 7 (Aim 3b)</b>					
<b>Subtask 1:</b> 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	
<b>Milestone(s) Achieved:</b> Identified mechanistically relevant molecules associated with ITH and endocrine resistance					

## B) What was accomplished under these goals? (Site 2: Data modeling and analysis)

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 have occurred in progress 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 due 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 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)

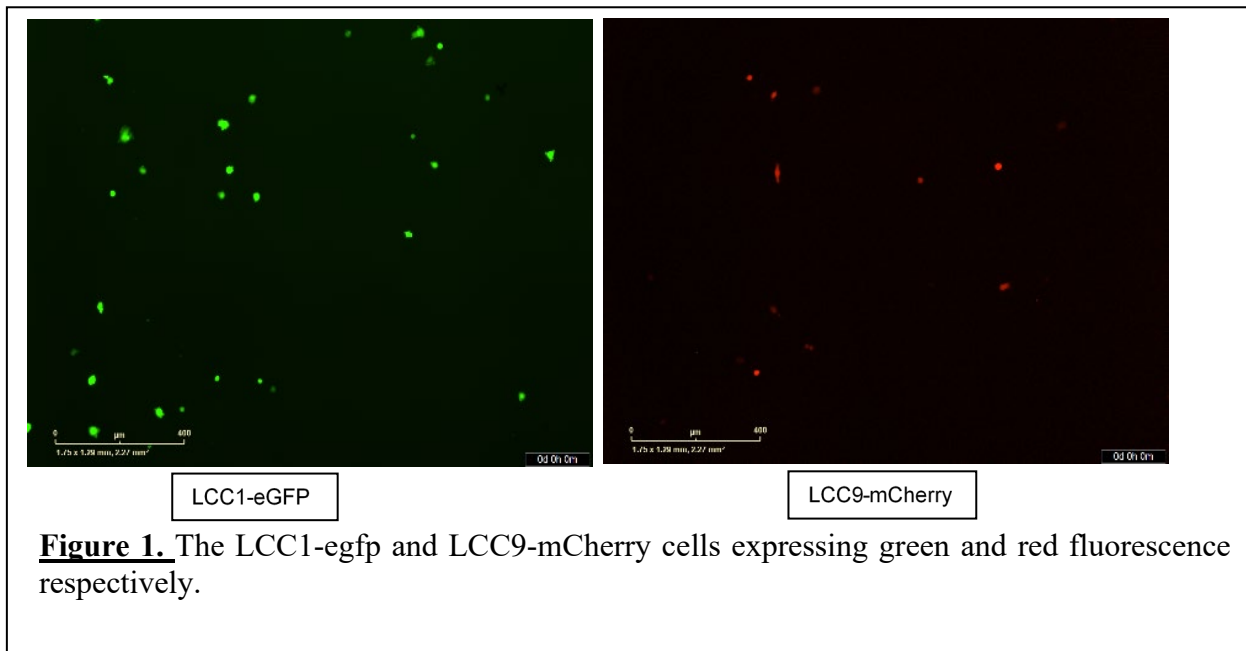
R+S cells at a single ration ± ICI (15+15=30)

Total = 90/experiment

We may do 3 such experiments over the 18-month period for 270 mice.

**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. Representative photographs (Figure 1) from these recent studies are shown, where LCC1 cells are labeled with enhanced green Fluorescence protein (eGFP) and LCC9 cells are labeled with red fluorescence protein, mCherry.



### **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)**

#### **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.

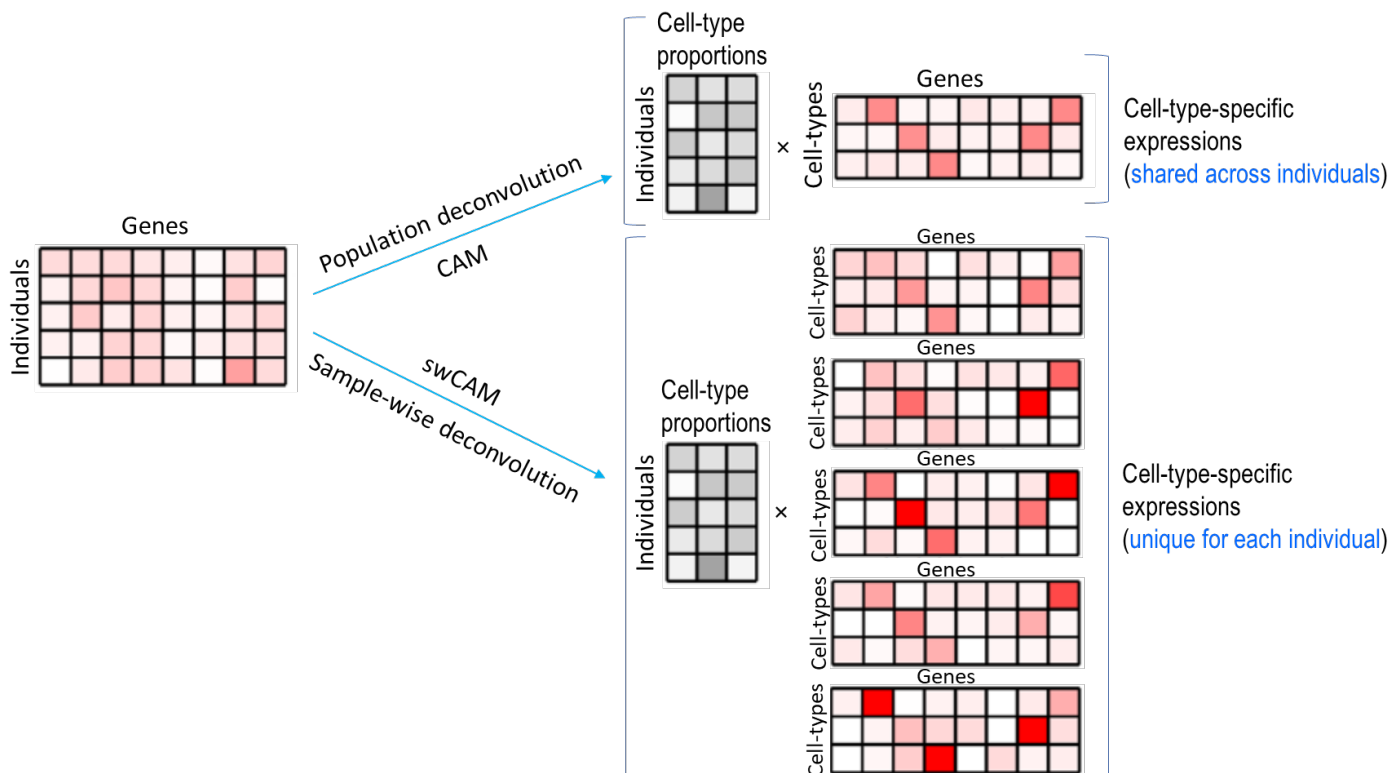
The R Scripts of Cosbin is freely available at <https://github.com/MinjieShen/Cosbin>

#### **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 1**). 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.



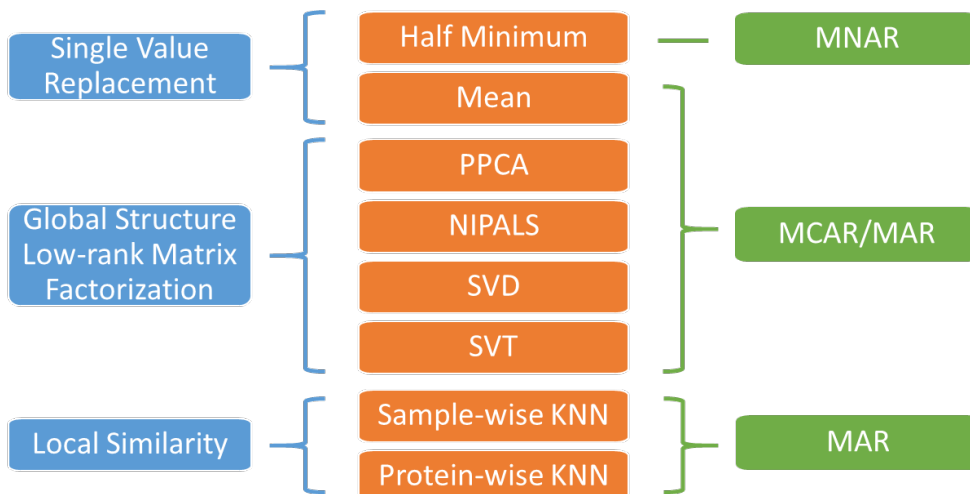
**Figure 1.** swCAM versus earlier CAM decomposition approach. Given bulk gene expression data from a heterogeneous tissue, earlier CAM approach aims at estimating a matrix of the cell-type proportions of the individuals and a matrix of the cell-type-specific transcriptome in the sample (shared across 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>

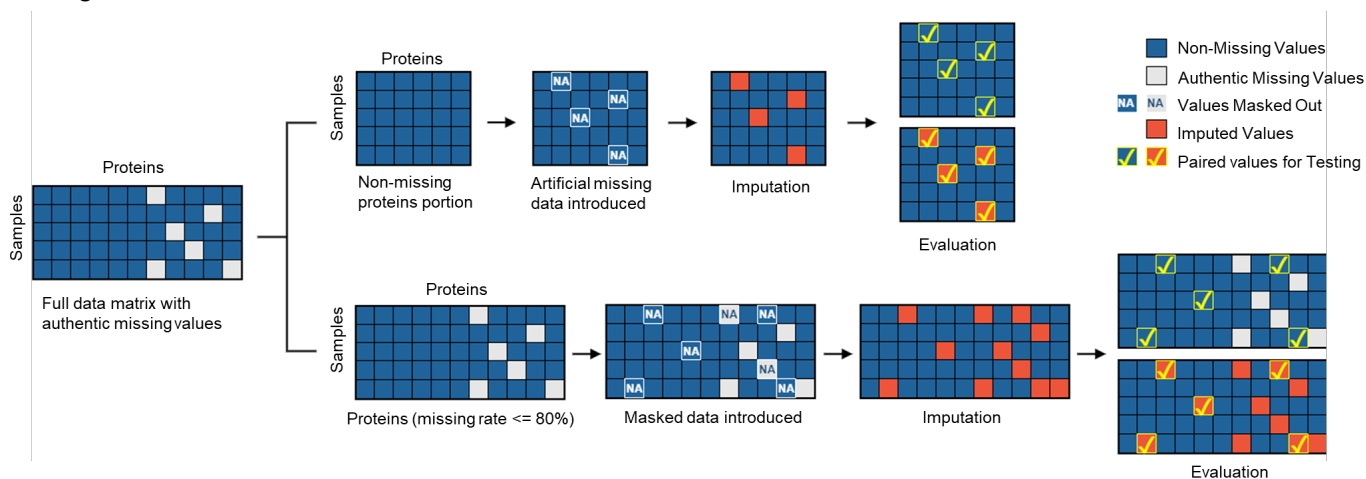
### B-3. Comparative assessment and novel strategy on methods for imputing proteomics data

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 2**). These methods are compared on both simulated and masked missing values embedded within real proteomics datasets, and performance is

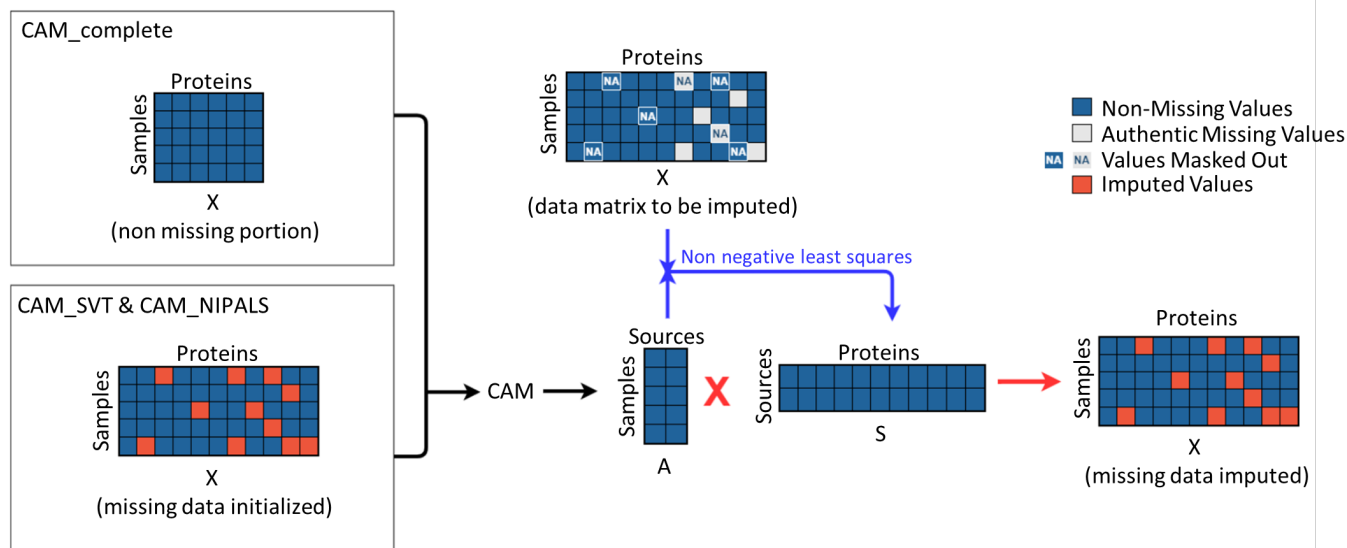
evaluated using three quantitative measures (**Figure 3**). We then introduce fused regularization matrix factorization, a low-rank global matrix factorization framework, capable of integrating local similarity derived from additional data types. We also explore a biologically-inspired latent variable 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 4**).



**Figure 2.** Comparative assessment of eight representative missing value imputation methods, divided into three categories.



**Figure 3.** Two-phased workflow of realistic simulation-based assessment on missing value imputation methods.



**Figure 4.** Workflow of the CAM based imputation method with two variant algorithms.

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

### **Major Task 7**

**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) What opportunities for training and professional development has the project provided?**

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 were the results 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) What do you plan to do during the next reporting period to accomplish the 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**

Describe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:

### **What was the impact on the development of the principal 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 tissues as the reviewers pointed out, there are some other 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 will not have reference either. 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.

### **What was the impact on other disciplines?**

Three software tools, i.e., Cosbin, swCAM, ProImput, can have broad applications in data sciences, engineering, and medical imaging.

### **What was the impact on technology transfer?**

Nothing to Report.

### **What was the impact on society beyond science and technology?**

Nothing to Report.

**Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of science, engineering, and the academic world on areas such as:**

## **5. Changes/Problems**

Nothing to Report.

### **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

### Journal 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

### 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

Software: Cosbin, swCAM, ProImput, DDN.

## 7. Participants & Other Collaborating Organizations

**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 and data interpretation

**Name:** Surojeet Sengupta

**Project Role:** Co-Investigator

**Nearest person months worked:** 4.2

**Contribution to project:** Re-established protocols and experimental conditions at the UMN

**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**