AD_____

Award Number: W81XWH-12-1-0107

TITLE: Use of a Novel Embryonic Mammary Stem Cell Gene Signature to Improve Human Breast Cancer Diagnostics and Therapeutic Decision Making

PRINCIPAL INVESTIGATOR: Charles M. Perou

CONTRACTING ORGANIZATION: The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599

REPORT DATE: October 2014

TYPE OF REPORT: Annual

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2014	Annual	30 Sep 2013 – 29 Sep 2014

4. TITLE AND SUBTITLE
Use of a Novel Embryonic Mammary Stem Cell Gene Signature to Improve Human Breast Cancer Diagnostics and Therapeutic Decision Making

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-1-0107

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Wahl, Geoffrey M.; Perou, Charles; Spike, Benjamin; Lasken, Roger

E-Mail: cperou@med.unc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

The University of North Carolina at Chapel Hill
Lineberger Comprehensive Cancer Center
450 West Drive, CB# 7295
Chapel Hill NC 27599-7295

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel 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
Our major goals are to determine whether Fetal Mammary Stem Cell (fMaSC) signatures correlate with response to chemotherapy and metastasis in different breast cancer intrinsic subtypes (AIM1), and to develop single cell sequencing to produce highly refined fMaSC signatures (AIM2). Accomplishing these aims will enable us to: 1) better categorize distinct cell types within the fMaSC population, 2) identify biomarkers for prospective stem cell purification and in situ localization, and 3) identify candidate stem cell regulatory pathways that should reveal therapeutic targets and improved prognosticators and response biomarkers. In the most recent funding period, our bioinformatic analysis identified subsets of fMaSC signature genes that are coordinately expressed in archived human breast cancer gene expression data sets and assessed their prognostic and/or predictive power. We have thus far identified one subset exhibiting significant prognostic value distinct from existing and commonly used clinical variables in the preliminary data sets we have analyzed. We have also adapted a new microfluidics-based, single-cell capture and library preparation system to improve reproducibility in the generation of gene expression profiles from individual fMaSC. These advances provide proof of the principles underlying this grant and leave us well positioned to achieve its aims.

15. SUBJECT TERMS
Breast Cancer Prognosis, Mammary Stem Cells, Embryonic Development, Single Cell Transcriptomics

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	10	USAMRMC

Table of Contents

Progress report for DOD/USAMRAA

USE OF A NOVEL EMBRYONIC MAMMARY STEM CELL GENE SIGNATURE TO IMPROVE HUMAN BREAST CANCER DIAGNOSTICS AND THERAPEUTIC DECISION MAKING

W81XWH-12-1-0106
The Salk Institute for Biological Studies
Geoffrey M. Wahl, Initiating PI

and

W81XWH-12-1-0107
University of North Carolina at Chapel Hill
Charles M. Perou, Partnering PI

Keywords.

breast cancer, stem cells, genomics, gene expression, fetal

Introduction.

This Idea Award Expansion proposes two aims to capitalize on discoveries made under the originally funded Idea Award, which included our identification and profiling of Fetal Mammary Stem Cells (fMaSCs) and uncovering of molecular similarities between fMaSCs and human breast cancers.

The first of these expansion aims (Aim1), proposes to refine the fMaSC gene expression signatures to better inform breast cancer disease modeling and design of prognostic and predictive metrics for chemotherapy response or metastasis. To this end, we have derived new RNA-Seq based fMaSC profiles that corroborate our earlier array based signatures but are more comprehensive and quantitative owing to technological improvements associated with sequencing (Dravis, Spike and Wahl in preparation). Further, in the course of our recently published collaborative study examining links between human breast cancers and mouse mammary tumor models, we identified unique enrichment patterns for the fMaSC signature and and specific mouse models (Pfefferle 2013). We have now discovered correlations between distinct fMaSC signature components and pathological complete response (pCR) to neoadjuvant chemotherapy in human breast cancer patients (Pfefferle 2014, manuscript in preparation), thus fulfilling one of the major goals of this proposal.

The second (Aim2), which comprises the major focus in the Wahl lab, took on the significant challenge of using single cell RNA-sequencing to deconstruct the fMaSC population into its component cell types. The purpose of this analysis has been to determine whether genes associated with breast cancer enriched signatures and/or

distinct mammary lineages (i.e. basal and luminal cells) are co-expressed in individual fMaSC cells denoting their uncommitted, multi-potent state, or alternatively, whether these complex signatures are in fact present only within different cells. These studies are simultaneously aimed at creating precise signatures for the fMaSC state, free from contaminating non-stem cell contributions. We also seek to identify candidate fMaSC biomarkers of potential relevance for mammary tumorigenesis and progression.

Overall Progress Summary.

Progress for Aim 1

As mentioned in our last Progress Report, the Perou lab is focused on Aim 1, which includes many computational analyses of existing gene expression databases in order to explore the prognostic and predictive potential of the fMaSC signature. This collaboration between the Perou and Wahl labs has already resulted in a publication focused on the classification of mouse models (Pfefferle, Herschkowitz et al. 2013), which now provides us with a rich database of mouse tumor profiles for investigations of fMaSC features in mouse and human tumors. We are currently writing a second manuscript focused on a detailed analysis of the fMaSC profile with our most recent finding being that the fMaSC signature is the most enriched in the murine WapINT3 and Class14 groups, both of which show stem cell features (Pfefferle et al., 2013); more specifically, in a new manuscript we have used transcriptomic profiles of fluorescence-activated cell sorting (FACS) fractionated normal mammary epithelial cell types from several independent human and murine studies to derive consensus gene cell type gene signatures to relate tumors to normal mammary cell types. Most human and murine tumor subtypes shared some but not all features with a specific FACS purified normal cell type.

As originally proposed, we have also been reanalyzing the original fMaSC genomic data to "refine" the fMaSC signature. Here, "refinement" means; 1) biological dissection of the fMaSC signature into sub-signatures, and 2) gene set reduction for translation to other technologies. Using a newly derived fMaSC signature coming from a supervised analysis of the fMaSC FACS fraction versus the fStromal+adultMaSC FACS fractions, we identified genes whose high expression better defines fMaSCs as a class of cells. Next we used this ~400 gene set to cluster 300 human breast tumors and determined that the fMaSC signature actually splits into 3 different sub-clusters; one sub-cluster is highest in basal-like tumors, another is highest in luminal tumors, and a third shows no subtype association. This ability to subdivide the fMaSC gene set hints at fMaSC multi-cellular differentiation potential since this single original signature can be broken into distinct smaller signatures that track different cell lineages.

We next explored the clinical potential of the three fMaSC sub-signatures using >500 tumors taken from the public domain, including ~100 ISPY samples as we had originally proposed. These tumors all came from patients treated with anthracycline and taxane containing neoadjuvant chemotherapy regimens. Even after accounting for the usual clinical and genomic features that have been used to predict the

likelihood of pCR, the complete fMaSC signature proved to be a significant response predictor. In addition, each of the fMaSC sub-signatures proved to be significant predictors; the fMaSC-basal-enriched signature predicted chemo-sensitivity, while the fMaSC-luminal-enriched signature predicted chemo-resistance. While we are encouraged by these results, we caution that they need validation using additional data sets, which we will do during the granted extension period.

During the extension period, we will also test the refined and full fMaSC signatures against the single cell data coming from Aim 2 from the Wahl lab (see below). Here, we want to determine if individual cells show the complete fMaSC signature, or whether individual cells are enriched for just one of the refined signatures. Here, the key question is whether individual single cells truly show both basal and luminal features, or whether these different features are restricted to different cells in the population; the data below suggests that both scenarios can occur.

Progress on Aim 2.

Obtaining RNA-seq data and reliable expression values from single cells (Aim 2a) has been fraught with widely recognized technical challenges from cell isolation to library preparation, and bioinformatics challenges for distinguishing technical noise from real biological variations. We have devoted significant effort in the initial year of this expansion award toward adaptation and development of RNA Sequencing and analytical approaches that achieve reproducible and reliable data. In year 2 we have now successfully sequenced hundreds of cells from multiple early stages of mammary development and adult mice and obtained interpretable data relating to individual stem cells and differentiating cells. To do this, we captured individual cells in separate nano-liter wells using the Fluidigm C1 microfluidic instrument and evaluated their viability microscopically. We then constructed libraries from viable cells using SmartSeq cDNA synthesis and Nextera XT chemistry prior to Illumina sequencing. We mapped sequence data from barcoded libraries containing 40-60

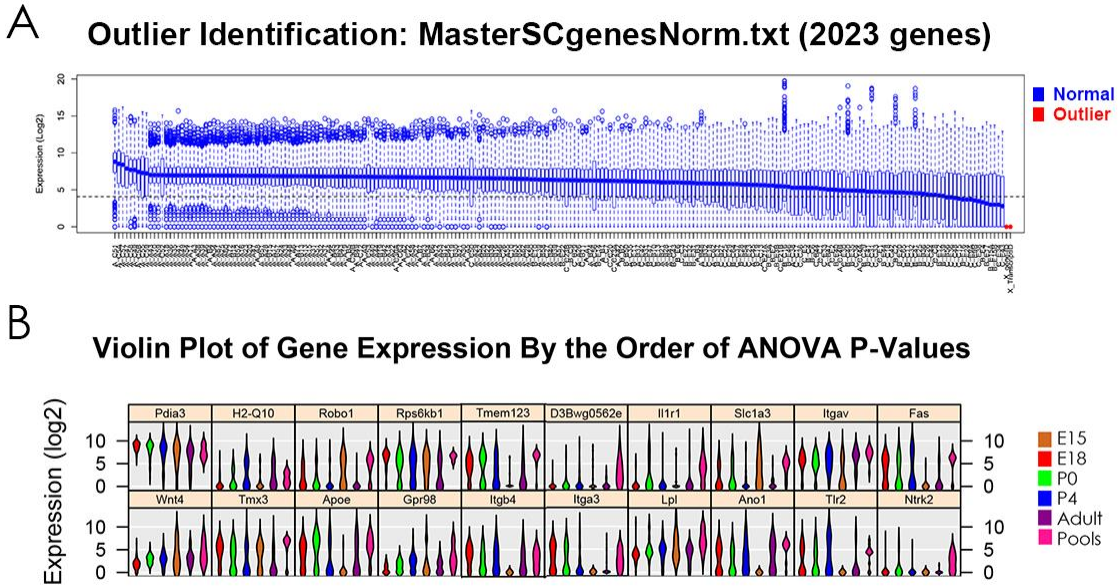

Figure 1. Hundreds of single cell samples pass a series of quality controls and yield differentially expressed genes associated with development and changing fMaSC number. A) Box and whisker plots (mean, 75th & 95th percentiles) indicating that our Single Cell RNA Seq libraries are robust, as the means and distributions across the samples indicate thousands of robustly expressed genes in each. B) Cell distributions (x axes) verses expression level for a representative set of genes indicate their differential expression at different developmental stages.

6

individual cells and pooled sample controls, and found that typical single mammary cells express 3000 – 6000 transcripts, which align well to gene exon models and match patterns seen from bulk preparation samples. To minimize technical variation and facilitate sample-to-sample comparisons, we employed numerous quality controls during library preparation and normalization strategies during analysis. The data in Figure 1a shows one such quality control measure aimed at identifying failed libraries as outliers. All samples in this analysis, apart from negative controls, robustly express genes from a set of over two thousand bearing high global expression (i.e. computationally summed and ranked), and are thus all suitable for downstream analysis.

We were able to identify the vast majority of expressed transcripts in cells from just 500,000 reads per sample indicating that our typical sequence depth of ~2.5 million reads/sample will provide high confidence gene calls. We developed a robust normalization strategy similar to Anders and Huber (2010) but restricting the algorithm to a subset of genes that includes many recognized housekeeping genes and avoiding normalization based on the most extremely overexpressed genes. Our normalization approach also avoids the assumptions of equal absolute transcripts per cell and equal RNA retrieval during lysis that are commonly made when normalizing on a "per million reads" basis (e.g. FPKM), or normalizing to "spike in" RNA transcript controls, respectively. Following these normalization steps we were able to identify many genes differentially expressed across the developmental stages collected, including those showing expected changes based on the literature and our own work (e.g. various integrins) but also many genes that have not been studied (Figure 1B).

While we are applying a variety of analytical methods to this data, we have found that the Monocle algorithm is useful for identifying differentially expressed genes and cell states (Trapnell, 2014). Unsupervised analysis of our single cell RNA-seq data using this algorithm organizes the data in a non-supervised fashion first in independent component space, and then according to differentiation status (Figure 2A). Differentiation status is inferred from a cell's position along a line connecting each sample in a 'minimum spanning tree'. This line, referred to as "pseudotime", is taken to reflect each cell's actual

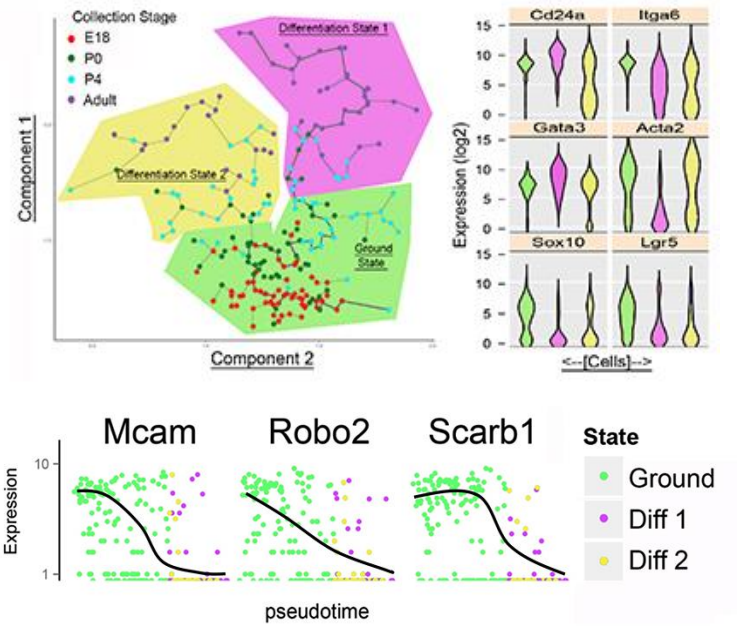

Figure 2. Single cell RNA-sequencing and Monocle analysis of fMaSCs. See text for description.

state of differentiation. The high quality of the RNA-seq data we can now routinely obtain using this technical and analytical pipleline is implied by the ability of this unsupervised analysis to clearly distinguish between two differentiated lineages (Differentiation State1 and 2) that directly correspond to luminal and myoepithelial cells based on their expression of known lineage correlated genes, such as Cd24a, Itga6, Gata3, and Acta2 (Figure 2B). Of particular note, this analysis has already satisfied one of our specified aims by revealing individual cells, particularly in the fMaSC/ground state set, that co-express lineage markers such as Keratin 14 and 8 (not shown), Gata3 and Smooth muscle actin (Acta2).

Additionally, other genes (e.g., Lgr5 and Sox10) appear to be expressed at high levels in the fMaSC/ground state set, and exhibit more restricted expression in most differentiated mammary epithelial cells and these are potential significance in defining the biology of the fMaSC state and of fMaSC-like cancers. For instance, we found that Sox10 is functionally relevant to the fMaSC state and Sox10 is highly expressed in numerous basal-like cancers (Dravis et al, Manuscript in preparation). Taken together, our observations indicate that Monocle will be useful for identifying other transcriptional regulators and pathways involved in lineage specification and acquisition of "stemness".

Thus, we are now deriving single cell fMaSC signatures (Aim 2B) and signatures for Monocle-defined cell states for enrichment analysis in archival breast cancer gene expression data sets using methodologies described under Aim 1. Gene expression comparisons between cell states predicted by the Monocle algorithm yielded ~4000 differentially expressed genes across all states tested. These genes are associated with diverse cellular compartments and gene ontology (GO) categories. For example, the Monocle "fMaSC/ground state" is distinguished from "differentiated lineages" by high expression of metabolic genes with potential roles in glycolytic metabolism (e.g., PKM2 embryonic splice isoform, and LDH) and genes regulating extracellular matrices (e.g., FN1, Col9a2, and Col4a2). This suggests roles for these genes in producing or maintaining the stem cell state, which we will test functionally using PKM2 and LDH inhibitors, and growth conditions that specifically employ matrices with fibronectin, and relevant collagens, etc.

Our confidence that this method can identify genes that actively contribute to the fMaSC state is bolstered by our studies with Lgr5 and Sox10, because we have found both to functionally contribute to fetal mammary stem cell biology and to enable significant fMaSC enrichment based on fluorescent reporters for each gene (C. Dravis, G. Wahl, in preparation; C. Trejo, G. Wahl, unpublished observation).

We are also mining our existing single-cell sequencing data, and will obtain additional data from pre-stem cell E15.5 rudiments to identify cell surface molecules expressed on candidate fMaSCs. We will look for genes that are downregulated during development, but retained in a subset of fMaSC-like adult "persistor" cells, and in fMaSC-signature enriched basal-like breast cancer cells. We expect this will be a rich discovery engine since we have already identified

interesting fMaSC cell surface candidates linked to breast cancer including: 1) Mcam (CD146), an epithelial-mesenchymal transition effector associated with TNBC; 2) Robo2, a growth factor receptor upregulated in inflammatory breast cancer; and 3) Scarb-1, a cell surface receptor for HDL, and a breast cancer risk factor (Figure 2C, each dot represents one cell; note high expression in fMaSC-enriched cells (green) and retained high expression rare adult cells (purple).

As we complete our analysis and validation of differentially expressed genes (Aim 2B) above, we are well positioned to complete the remainder of this aim by using these signatures and markers in tumor settings (Aim 2C) computationally as in Aim 1, or by detecting fMaSC-like cells *in vivo* in cancer mouse models and human tumor samples.

Conclusion.

We have used FAC sorting data coming from human and mouse adult mammary gland, and coming from the fetal mammary rudiment, to define gene expression profiles of different mammary epithelial cell populations. Using novel statistical analyses, we define gene expression that are characteristic for these different mammary cell types, including the normal mammary fetal stem cell, and then use these signatures to investigate the biology of human breast tumors. We find that the fetal mammary stem cell signature, when seen in human tumors, tends to predict an increased likelihood of achieving a pathological complete response to neoadjuvant chemotherapy. This finding is being further investigated, but if validated, shows the power of studying epithelial development for informing human tumor biology.

Publications, Abstracts and Presentations.

Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 2013 Nov 12;14(11):R125.

Inventions, Patents and Licenses.
none

Reportable Outcomes.

Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 2013 Nov 12;14(11):R125.

Other achievements.
none

References.

Anders and Huber, Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106

Dravis C, Spike BT, Trejo C, Harrell C, Perou CM, Wahl GM. On the role of Sox 10 in mammary stem cells. ***In Preparation.***

Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 2013 Nov 12;14(11):R125.

Pfefferle JM, Spike BT, Wahl GM, Perou CM. Luminal progenitor and fetal mammary stem cell expression features predict breast tumor response to neoadjuvant anthracycline and taxane chemotherapy. ***In Preparation.***

Trapnell et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381-6