AD________________

Award Number: W81XWH-11-1-0402

TITLE: Mammary Cancer and Activation of Transposable Elements

PRINCIPAL INVESTIGATOR: Dr. John Edwards

CONTRACTING ORGANIZATION: Washington University
Saint Louis, MO 63130-4862

REPORT DATE: September 2014

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution 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

1. REPORT DATE September 2014	**2. REPORT TYPE** Annual	**3. DATES COVERED** 1 Sep 2013 – 31 Aug 2014

4. TITLE AND SUBTITLE
Mammary Cancer and Activation of Transposable Elements

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0402

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John R. Edwards, PhD

E-Mail: jedwards@dom.wustl.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Washington University
1 Brookings Drive
Saint Louis, MO 63130

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

The purpose of this project is to investigate molecular events occurring in the preclinical stages of mammary cancer. Specifically, the project investigates the intersection between the development of genome demethylation, retrotransposon transcriptional activity, and retrotransposon-driven transcription of cellular genes in an engineered mouse model of mammary cancer. During the last 12 months, my collaborator Dr. Peaston completed the mouse breeding for the project, and collecting material for planned molecular analyses is now complete. We have received DNA for all replicates. We have made and performed initial of sequenced methylation profiling libraries for the first replicate and are in the process of making libraries for the others. and are currently making methylation profiling libraries. We have continued to make improvements to our methodologies for integrating and interpreting genome-wide expression and methylation data that will be used for the analyses in this project. We believe that the eventual findings will provide insights into understanding the role of genome hypomethylation and expression of retrotransposons in cancer ontogeny, and may impact cancer prevention in the future.

15. SUBJECT TERMS

Breast cancer, epigenetic, DNA methylation, retrotransposon, preclinical cancer development, deep sequencing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	8	

Table of Contents

Introduction

This project is designed to address the subject of mammary cancer development. The purpose of the project is to investigate molecular events occurring in the preclinical stages of mammary cancer; the results may lead to insights into cancer prevention in the future. Specifically, the project investigates the intersection between genome demethylation, retrotransposon transcriptional activity, and retrotransposon-driven transcription of cellular genes. Retrotransposon promoters are well recognized to function as alternative promoters for different cellular genes, generating chimeric transcripts that may or may not function in the same way as transcripts from the regular gene promoter. Transcriptional activation of retrotransposons is strongly linked with their CpG DNA methylation, and global genomic demethylation is one of the commonest molecular changes in malignancies. The project tests the hypothesis that, in preclinical stages of tumour development, progressive genomic demethylation leads to increased transcriptional activity of retrotransposons and this, in turn, leads to transcription of otherwise silent genes, potentially setting up molecular conditions that favour cancer development. Our collaborator, Dr. Anne Peaston, developed a genetically engineered mouse model in which a specific mammary cell population is fluorescently marked upon initial transcriptional activation of the SV40 large T antigen (SV40Tag) oncogene. SV40Tag is transcriptionally activated during pregnancy and lactation, and the mice are predisposed to develop mammary cancer after 3 pregnancies and lactations. Using this model, populations of marked cells can be collected for integrated analysis of gene expression, promoter usage, and DNA methylation after defined amounts of exposure to SV40Tag during different stages of preclinical cancer development.

Body

The relevant sections from the Statement of Work are shown in the table below with corresponding goals and results. A no cost extension was received to continue this work until December 2014 in light of some of the technical issues with the mouse breeding experiments (see Dr. Peaston's, collaborating PI, 2013 report for more information). While waiting on the shipments of DNA samples to begin processing them, we have worked on methodological improvements to streamline sample preparation and began to establish an analysis pipeline that can handle the three data types that will be produced in this proposal. This pipeline will greatly facilitate final analysis and interpretation of results for this project in the upcoming year. These advancements and their importance to the project are described below. In particular we feel confident that the streamlined protocol and analysis tools for Methyl-MAPS that we have developed will easily allow us to complete all Methyl-MAPS analyses and final computational analyses outlined in the Statement of Work before the project end.

Year 1 & 2: Items from Statement of Work Relevant to Edwards Lab.

Year, Months	Goal		Result
Y1 1-3 4-6 7-9 10-12 Y2 1-3 4-6 7-9 10-12	4. 9. 5. 7. 8. 8. 9,10,11. 8.	• Set up schedule for formal monthly electronic lab meeting between Peaston lab and Edwards lab. And regularly hold meetings.	• An informal schedule was set up with a plan for regular meetings
Y1 4-6	6.	• Preliminary Methyl-MAPS analysis of pilot virgin samples	• Initial libraries constructed and initial sequencing has been performed. Deeper sequencing is underway.
Y1 10-12	7.	• Methyl-MAPS library preparation and sequencing for replicate #1 uniparous & triparous control and tumor-prone	• Initial libraries constructed and initial sequencing has been performed. Deeper sequencing is underway.

Y2 1-3	6. 7.	• Preliminary Methyl-MAPS analysis of replicate #1 • Review DNA loci of interest for bisulfite sequencing in light of PCR results and library analysis, Continue bisulfite sequencing assessment of DNA loci of interest	• Initial library QC was performed and everything looks good. Deeper sequencing results are now being obtained. Once ready a full analysis will be performed. RNA-seq data has been received from Dr. Peaston and is currently being analyzed.
Y2 4-6	3. 5.	• Methyl-MAPS library preparation and sequencing for replicate #2 uniparous & triparous control and tumor-prone • Library analyses replicate #2 and preliminary comparisons with replicate #1	• Material has been received and libraries are nearly constructed.
Y2 7-9	4.	• Methyl-MAPS library preparation and sequencing for replicate #3 uniparous & triparous control and tumor-prone	• Material has been received. There are some concerns that there is sufficient material available for the analysis. After the second replicate is complete a decision will be made whether to pursue a third or use this material for subsequent validation of the genomic results.
Y2 10-12	2. 3. 5.	• Finalize data analysis • Prepare a report for publication of the results	• Analysis pipelines have been developed and are in place for when data is generated. Reports will be finalized as the data is.

Methyl-MAPS Analysis replicates 1-3

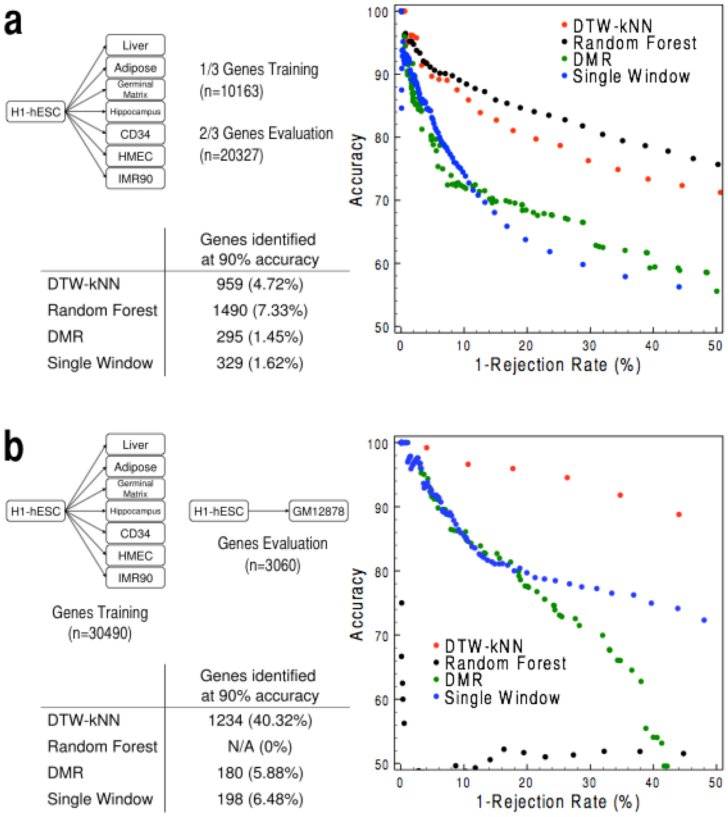
DNAs have been received for all replicates. One “replicate” consists of four samples, or each combination of uniparous or triparous, and control or SV40+ mice. Methyl-MAPS libraries have been constructed for each sample in the first replicate. Initial low-pass sequencing of these libraries has been performed for quality control. 2.5-6.5 million tags for each McrBC or RE library from each sample were mapped to the mouse mm10 genome for an average coverage that ranged from 2.5-4.5x. Additional sequencing to raise the coverage is underway.

For replicates 2 and 3 quality control analysis was performed on DNA from each sample received from the University of Adelaide. Methyl-MAPS libraries are nearly completed for replicate 2. There is some concern whether there is sufficient material for replicate 3. If there is not, then the DNA will be used for validation of target regions identified in the first two replicates.

Computational Pipeline Improvements

We have developed the WIMSi (Washington University Methylation Signatures) pipeline to enable us to use genome-wide methylation and expression data to examine the relationship between DNA methylation and expression. An initial description of the method and results can be found in VanderKraats et al, 2013[1].

In brief, current computational tools focus on methods to compute accurate methylation levels from the data, methods to determine differentially methylated regions, and visualization tools, such as genome browsers. These approaches have elucidated the genomic organization of these marks, but they do not sufficiently address how changes at individual loci potentially affect function. Correspondingly, such methods find only a modest negative correlation between differential DNA methylation at promoters and expression. While it is possible that DNA methylation is not a strong modifier of expression, it is more likely that our computational tools are insufficient. We hypothesize that stronger associations are not observed because existing analysis methods oversimplify their representation of the data such that they do not capture the diversity of existing methylation patterns. This includes changes at the CpG island at a gene’s TSS, changes at CpG island shores and the formation of long partially-hypomethylated domains. In addition, there are not methods to systematically search for new patterns.



	Genes identified at 90% accuracy
DTW-kNN	959 (4.72%)
Random Forest	1490 (7.33%)
DMR	295 (1.45%)
Single Window	329 (1.62%)

	Genes identified at 90% accuracy
DTW-kNN	1234 (40.32%)
Random Forest	N/A (0%)
DMR	180 (5.88%)
Single Window	198 (6.48%)

Figure 1 Comparison of an adaptation of our approach (DTW-kNN, dynamic time warping-based k-nearest neighbors) in VanderKraats et al 2013, other methods from literature (DMR and Single Window), and a newly developed Random Forest approach based on methods from the analysis of chromatin signals. Performance is plotted as one minus the rejection rate, (equivalent to the percentage of genes returned in the evaluation set) versus the accuracy (correct prediction of the direction of expression change). An accuracy of 50% is equivalent to a random guess, and the top right corner indicates optimal performance. Training and evaluation data are from whole genome bisulfite sequencing and RNA-seq from the Human Epigenome Atlas Project, Hon et. al *Genome Res.* 2012, and ENCODE. **(a)** Random Forest and DTW-kNN compare favorably when evaluated on genes from the same samples used to create the training model. **(b)** However, DTW-kNN outperforms all methods in the paradigm where a training model is built on one set of samples and evaluated on an independent dataset. DMR and single window approaches were tuned using a grid-search of published parameters.

We have developed a new set of tools for discovering differential methylation patterns associated with expression change using genome-wide high-resolution methylation data: we represent differential methylation as an interpolated curve, or signature, and then identify groups of genes with similarly-shaped signatures and corresponding expression changes. Our methods uncover a diverse set of DNA methylation patterns that are conserved across a variety of normal and cancerous tissues and cell lines. Source code from the method is publically available for download at: sourceforge.net/projects/wimsi. These approaches have been refined and applied to study methylation changes in AML after treatment with DNMT inhibitors (Lund et al. 2014) and to understand methylation signatures in senescent cells that resemble those found in tumors (Cruickshanks et al 2013).

Gene List Generation

Enumerating a list of genes for which expression and methylation changes are potentially linked is a primary interest of any genome-wide methylation profiling experiment. The WIMSi method outlined above is tuned to discover patterns, and thus produces a conservative gene list potentially prone to false positives. Our initial publication describes one way to use this method to generate a gene list, however, we have further refined our approach using a set of supervised machine learning methods.

At one level, the basic principal behind WIMSi is to define the distance between any two methylation signatures in terms of shape similarity or distance between curves. Once these distances are defined, we use a k-nearest neighbors based algorithm in which the expression labels for one set of genes is known and the expression of a new gene is predicted based on the similarity of it's methylation pattern. For example if a new gene has a methylation signature similar to 10 (or k =10) other genes whose expression is silenced then likely the new gene is also silenced. Our initial results have shown that this is a powerful and stable method for using methylation patterns to predict expression. This method works much better than classical DMR algorithms at predicting expression (Fig. 1) and thus at finding associations between methylation and expression. Our initial results also show this method to be comparatively robust to choices of parameters across datasets. For instance, in our method if we train a predictive model on one dataset we can use this model to accurately predict expression changes in another. Whereas it appears these parameters must be re-optimized for every dataset with traditional DMR-based methods[2].

Our findings suggest that the role of DNA methylation cannot be fully described by simply characterizing every gene as "methylated" or "unmethylated". Using our new method, we have found and described a variety of methylation patterns that correlate with expression change. The true power of this method is in its ability to discover and separate distinct patterns without *a priori* knowledge about existing correlations, which cannot be accomplished with contemporary approaches. This allows us to realize the full potential of unbiased genome-wide profiling of DNA methylation to reveal previously unknown information about methylation's functional role. This ability will be especially important when examining regulatory elements within retroelements as will be performed in this work.

One additional implication of these results also becomes clear. The simplified models used in prior approaches at best produce weak correlations. However, if one considers a more formal description of the underlying patterns of methylation changes, methylation and expression data are highly correlated. This tool was designed to start from a list of expression data, corresponding transcription start sites (TSSs) and high-resolution genome-wide methylation data such as from Methyl-MAPS. Thus it fits perfectly into the framework of this proposal where we will have expression and Methyl-MAPS methylation data for each sample. The adaptations to accommodate these datasets to address the regulation of retrotransposons are straightforward and we have performed some initial analyses of LTR-driven lncRNAs as a successful proof-of-concept. We have an established pipeline in place and ready for the data as it is produced in the upcoming year.

Key Research Accomplishments

- Continued development of new computational tool to combine genome-wide expression and methylation data to output a list of genes where methylation likely contributes to their silencing or activation.
- Construction and initial sequencing of methylation profiling libraries from the first experimental replicate.

Reportable Outcomes

Manuscripts

Lund K, Cole JJ, VanderKraats ND, McBryan T, Pchelintsev NA, Clark W, Copland M, Edwards JR, Adams PD. DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biol,* 2014; 30;15(8):406.

Cruickshanks HA, McBryan T, Shah PP, Nelson DM, Donahue G, VanderKraats ND, Edwards JR, Berger SL, Adams PD. Features of the cancer epigenome are acquired as primary human cells approach senescence. *Nature Cell Biology.* 2013; 15:1495-1506.

Abstracts

Schlosberg CE, VanderKraats ND, Hiken JF, Weinberger KQ, Ju T, Edwards JR. (2014) "Modeling complex patterns of differential DNA methylation that strongly associate with gene expression changes." *22nd Annual International Conference on Intelligent Systems for Molecular Biology*, July 13-15.

Conclusions

The streamlined Methyl-MAPS protocol and the computational analysis pipeline we have now established will be invaluable for pushing the project ahead. The primary tasks for my lab were to provide Methyl-MAPS genome-wide methylation profiling for tumor DNA from the uniparous and triparous female mice from Dr. Peaston's mouse model, and to perform integrative analysis of methylation and expression data. Methyl-MAPS methylation analysis will be finished shortly. The computational tools we have developed are designed to work with annotated genes as we have outlined, but can also be expanded to any transcriptional unit with a known TSS and known expression value. We will thus be able to integrate each of the datasets generated in this project to address the hypothesis that, in preclinical stages of tumor development, genomic demethylation leads to increased transcriptional activity of retrotransposons and this, in turn, leads to transcription of otherwise silent genes, potentially setting up molecular conditions that favor cancer development.

References

1 Vanderkraats, N. D., Hiken, J. F., Decker, K. F. & Edwards, J. R. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res* **41**, 6816-6827, doi:10.1093/nar/gkt482 (2013).

2 Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology* **13**, R83, doi:10.1186/gb-2012-13-10-r83 (2012).