REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

The 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 the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 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 oenalty 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.

1. REPORT DATE (DD-MM-YYYY) 30-08-2016	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1-Jul-2014 - 31-Mar-2015

4. TITLE AND SUBTITLE Final Report: Reconstructing cell lineages from single-cell gene expression data: a pilot study	5a. CONTRACT NUMBER W911NF-14-1-0360
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 611102
6. AUTHORS Guo-Cheng Yuan	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Dana-Farber Cancer Institute, Inc. Biostats & CompBiology 450 Brookline Avenue Boston, MA 02215 -5450	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSOR/MONITOR'S ACRONYM(S) ARO
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) 65812-MA-II.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

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

14. ABSTRACT

The goal of this pilot study is to develop novel mathematical methods, by leveraging tools developed in the bifurcation theory, to infer the underlying cell-state dynamics from single-cell gene expression data. Our proposed method contains two steps. The first step is to reconstruct the temporal order of the cells from gene expression data, whereas the second step is to model the dynamic changes of gene expression patterns. The tools developed here will be useful for reconstruct developmental trajectories from single-cell gene expression data.

15. SUBJECT TERMS

Mathematics; Computational Biology; Bifurcation; Development; SIngle-Cell Genomics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Guo-Cheng Yuan
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-582-8532

Report Title

Final Report: Reconstructing cell lineages from single-cell gene expression data: a pilot study

ABSTRACT

The goal of this pilot study is to develop novel mathematical methods, by leveraging tools developed in the bifurcation theory, to infer the underlying cell-state dynamics from single-cell gene expression data. Our proposed method contains two steps. The first step is to reconstruct the temporal order of the cells from gene expression data, whereas the second step is to model the dynamic changes of gene expression patterns. The tools developed here will be useful for reconstruct developmental trajectories from single-cell gene expression data.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received	Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Eugenio Marco	0.60
FTE Equivalent:	**0.60**
Total Number:	**1**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Guo-Cheng Yuan	0.05	
FTE Equivalent:	**0.05**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

Title: Reconstructing cell lineages from single-cell gene expression data: a pilot study

(W911NF-14-1-0360, PI: Guo-Cheng Yuan)

Funding period: 01 July 2014 – 31 March 2015

A. Study Goal

The goal of the short-term research proposal is to develop and test a computational method for single-cell analysis in order to assess the feasibility of a recently submitted research proposal entitled "Single cell analysis of dedifferentiation and transdifferentiation in mammalian regeneration" (hereafter referred to as the Main Proposal) in collaboration with Prof. Ken Muneoka.

B. Studies and Results

In the past year, we have continued to systematically investigate the targeting mechanisms of epigenetic factors using a two-step approach. In the first step, we investigated the cell-type dependent plasticity of epigenetic patterns and the role of DNA sequence in mediating the degree of plasticity. In the second step, we further incorporated gene expression data to identify transcription factors that play a role in modulating the epigenetic patterns in a cell-type specific manner. To this end, we developed and experimentally validated a computational method to predict the genome-wide distribution of epigenetic plasticity and cell-type specific recruiting factors [1]. In addition, we also applied computational and experimental approaches to study a number of other related issues, such as the stability of bivalent domains. Our work has resulted five manuscripts that have been or are close to be published in peer-reviewed scientific journals [1-5]. In the following, I will describe the progress we have made during this funding period.

Aim 1: We will develop a computational approach to investigate the role of DNA sequences in the regulation of genome-wide epigenetic patterns

For model development, we chose H3K27me3 as a proof-of-concept due to its well-recognized important role in developmental control and the availability of large amount of public data. We obtained the H3K27me3 ChIP-seq datasets in 19 human cell lines from the ENCODE consortium and quantified the plasticity of H3K27me3 occupancy across these cell-types at each genomic locus. We then focused on the highly plastic regions (HPRs) because they are directly related to cell-type identity maintenance, as suggested by function enrichment and sequence conservation analyses. Consistent with the literature, we found that the majority of HPRs are proximal to CpG islands, which were previously shown by Brad Bernstein and others to play an important role in recruitment of Polycomb group (PcG) proteins. On the other hand, 44% of the HPRs are located in distal regions, where the recruitment mechanism is less clear.

To investigate the role of DNA sequence in mediating the degree of plasticity, we applied our N-score model to predict the location of the HPRs based on DNA sequence information. Our model provides substantial prediction power (AUC = 0.82), suggesting the DNA sequence indeed plays a significant role. Of note, many of the proximal HPRs can be predicted by using GC- and CpG-content alone, but additional sequence features are needed for accurate prediction of distal HPRs. We also investigated the mechanisms for establishment of cell-type specific patterns at the distal HPRs. The results are described under Aims 2 and 3.

Recent studies suggest another important mechanism for PcG recruitment which is through interaction with lincRNAs. By using RIP-chip experiments, John Rinn's lab identified about hundreds of lincRNAs that may interact with EZH2/PcG. To investigate a role of the DNA sequence in mediating PcG-lincRNA interactions, we adapted a BART model to predict the EZH2-bound lincRNAs and obtained excellent prediction accuracy (AUC = 0.92). Our work also provides new insights into the RNA structure in mediating such interactions. A manuscript is currently in preparation.

Lastly, we were invited by Dr. Suzanna Vinga to contribute a review paper in a special issue of Briefings of Bioinformatics [5]. In this paper, we surveyed recent development of alignment-free methods and their applications in epigenomics. Our survey also suggests the availability of unexplored methods that could lead to new discoveries in epigenomics research.

Aim 2: We will develop a computational approach to predict tissue-specific epigenetic patterns by integrating DNA sequence information together with gene expression data.

Following our initial analysis of the H3K27me3 plasticity, we aimed at identifying transcription factors that play a role in modulating cell-type specific patterns. To this end, we developed a computational pipeline that integrates both DNA sequence and gene expression data in four steps: 1) identification of the subset of HPRs that are specific to a particular cell-type; 2) identification of enriched TF motifs using the matched genomic background; 3) refining the list of motifs by testing the enrichment in center vs flanking regions; and 4) mapping the motifs to TFs and removing non-functional TFs by integrating gene expression data.

In total, our analysis predicted 41 cell-type specific TF-HPR associations in the 19 ENCODE cell lines. These predictions were validated by using independent data available in the literature, including ChIPseq data, shRNA data, and functional analysis. As such, we found significant support for most of our predictions. As a more stringent validation, we conducted experiments to directly test the function of predicted TFs in a model system, as described below.

Aim 3: We will experimentally validate the computational predictions.

We used primary human erythroid pregenitors (ProEs) as a model system to validate our predictions. In previous work [6], we generated genome-wide profiles of transcription, histone modification, and transcription factors in adult and fetal ProEs to investigate the mechanism underlying developmental stage-specific gene activities. We combined the H3K27me3 data in ProEs and in the ENCODE cell lines to carry out our computational pipeline.

Surprisingly, our model predicted TAL1, which is a principal regulator for hematopoietic development and commonly known as a transcriptional activator, to play a major role in mediating ProE-specific H3K27me3 patterns, suggesting a previously unrecognized function of TAL1. To identify the cofactors associated with this repression role, we searched for TF motifs that are differentially enriched between the TAL1+H3K27me3 and TAL1+H3K27ac regions, and predicted GFI1B as a leading candidate. To test these predictions, we conducted ChIPseq and co-IP experiments as validation. Our ChIPseq data confirm that TAL1 colocalizes with H3K27me3 at the ProE-specific HPRs, and our co-IP data show that EZH2/PRC2 can pull down both TAL1 and GFI1B, and that GFI1B can pull down EZH2. These data strongly validate our computational approach and highlight the utility of chromatin plasticity analysis in uncovering novel mechanisms.

We developed an allelic-imbalance approach for studying the molecular functions of GWAS variants [4]. As a proof-of-concept, we focused on a well-characterized enhancer of BCL11A, which was previously shown to repress HbF in adult erythroid cells. The region contains a number of genetic variants associated with the HbF level. To test the function of these variants, we adapted a TALEN-based assay to delete the target DNA sequence and identified a single variant that causally reduced GATA1 binding and increased HbF expression. This approach will be used to experimentally validate our computational predictions in the future.

C. Significance

We have developed a powerful, generally applicable approach for investigating the mechanism underlying the plasticity of epigenetic patterns. Our approach systematically investigates the role of DNA sequence in modulating plasticity, and provides a useful tool to identify key regulators modulating cell-type specific epigenetic patterns. Our approach has lead to new insights into the multi-faceted role of master regulators, such as TAL1, in the maintenance of cell identity.

D. Plans

We plan to extend our previous work by applying our model to other epigenetic marks. In this direction, we have initiated collaboration with Dr. Andrew Feinberg's group, who has pioneered cancer epigenetics and recently discovered the extensive DNA methylation variability among cancer patients. By using our computational pipeline, we are working to identify the underlying regulator for DNA methylation variability. We will also extend our model by further incorporating multiple epigenetic marks and investigating the plasticity of the combinatorial states. We are currently funded by NHGRI (R21HG006778) to develop a statistical model that characterizes the hierarchical chromatin structure by integrating multiple chromatin marks and will generate genome-wide maps of hierarchical, combinatorial chromatin states in ENCODE cell lines. These maps will serve as the basis for study combinatorial state plasticity. The predictions will be experimentally validated by using the ChIPseq, shRNA, and genome editing methods as we done before. Lastly, we have started to package our algorithm as an open-source, user-friendly software to make it accessible to the broad community.

E. Presentations

Invited Presentations (given by Guo-Cheng Yuan unless otherwise indicated)

- "Analysis of Gene Expression at the Single-Cell Level", Bioconductor Annual Meeting, Boston, Massachusetts, 07/31/2014
- "Mapping Cellular Hierarchy by Single-Cell Gene Expression Analysis", Single-Cell RNA Sequencing Workshop, Harvard Stem Cell Institute, Boston, 10/3/2014

Contributed Presentations

- "Bifurcation analysis of single-cell gene expression ", ISCB Annual Meeting/Single-Cell Genomics Special Interest Group, Boston, Massachusetts, 07/12/2014
- "Bifurcation Analysis of Single-cell Gene Expression", Single-Cell Genomics Meeting, Karolinska Institute, Stockholm, Sweden, September 9–11, 2014

F. Publications

1. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, Yuan GC. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014 Dec 30;111(52):E5643-50. PMCID:PMC4284553

G. Project-Generated Resources

The SCUBA software generated by the studies described in the above has been deposited in Github. URL: https://github.com/gcyuan/SCUBA