AWARD NUMBER: W81XWH-15-1-0243

TITLE: Development of a Prognostic Marker for Lung Cancer Using Analysis of Tumor Evolution

PRINCIPAL INVESTIGATOR: Edward F. Patz, Jr., M.D.

CONTRACTING ORGANIZATION: Duke University, Durham, NC

REPORT DATE: AUG 2018

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2018	Annual	1AUG2017 - 31JUL2018

4. TITLE AND SUBTITLE

Development of a Prognostic Marker for Lung Cancer Using Analysis of Tumor Evolution

5a. CONTRACT NUMBER
W81XWH-15-1-0243

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Edward F. Patz, Jr., M.D.

E-Mail: patz0002@mc.duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Duke University
Durham, NC 27705

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 goal of this project is to sequence the exomes of single tumor cells from tumors in order to construct evolutionary trees, the characteristics of which will be used to predict whether a tumor will metastasize or not. Since the last annual report, we continued to collect fresh lung tumor specimens, process them into single cell suspensions, and isolate and amplify DNA for sequencing. Laboratory Corporation of America (LabCorp) took over the sequencing for this project, and all sequencing is complete. We have received all the data and are in the midst of data processing, mutation analysis, and construction of phylogenetic trees.

15. SUBJECT TERMS

NSCLC; tumor evolution; whole exome sequencing

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

Table of Contents

Page

1. Introduction

There is currently no consistent way to determine how aggressive or indolent a lung tumor will be even among patients with the same radiographic findings, histology, stage, or molecular markers. The purpose of this grant is to apply evolutionary analytical methods developed to study expansion and migration of populations to tumor biology in order to produce a prognostic marker in cancer. As with the Darwinian evolution of populations, the evolution of tumor cells within a tumor can be diagrammed on a phylogenetic tree. The more diverse a tumor's phylogenetic tree, the more likely it is that there are cells within it that have acquired the genetic alterations that allow them to proliferate at an increased rate, migrate, and metastasize. We will develop and validate a novel, objective, and measurable "prognostic score" based on the probability that some tumors will be *aggressive* and metastasize, and other tumors will be *indolent* and not metastasize. We first will perform whole exome sequencing of individual tumor cells from the tumors of a training set of patients (half early stage, half late stage). We will reconstruct each tumor's phylogenetic tree (a map of the clonal evolution reflecting divergence and heterogeneity), and compare the tree patterns from early stage NSCLC (indolent tumors without metastasis) to those from late stage disease (tumors with metastasis). We will use a combination of *tree features* (including branch length and tree shape) to generate a prognostic score (a continuous variable and a measure of tumor heterogeneity) that separates tumors with very different phenotypes (indolent vs. aggressive). We will derive the prognostic score by determining the probability of each individual tumor's outcome in the pilot training study, and then validate this strategy in an independent set of patients. An accurate prognostic score could significantly change clinical management and improve outcomes.

2. Keywords

NSCLC; tumor evolution; whole exome sequencing

3. Accomplishments

Specific Aim I: Isolate individual tumor cells from 10 patients with stage I non-recurrent NSCLC and 10 patients with advanced stage NSCLC.

Major Task 1. Isolation of individual tumor cells by flow cytometry (60 cells from each of 20 patients, with 40 cells needed). To date we have collected 38 tumors, but were unable to isolate sufficient cells from all tumors, due to several issues including tumor size and cell necrosis. We have now isolated a sufficient number of single tumor cells for analysis from 16 patients. The patients whose tumors did contain sufficient isolatable cells are equally divided between non-recurrent and recurrent or metastatic cancers. Therefore, we decided to proceed with the project, sequence the single cells from each patient, and perform phylogenetic tree analysis. We continue to collect addition specimens if needed.

Specific Aim II. Perform single cell whole exome sequencing on 40 individual cells isolated from each tumor.

Major Task 2. Amplify genomic DNA from 820 cells (40 cells from each of 20 patients plus 1 germ line control from each), and

Major Task 3. Perform exome sequencing on 820 genomes (with subtasks to include sequencing runs, alignment of sequences to the genome, quality control filtering, variant calling, and resequencing selected variants for validation).

We continued to use our standard operating procedure for isolating single tumor cells and amplifying DNA from each cell for whole exome sequencing as described in previous reports. Suspensions that consisted primarily of cells between ~12-20 μm diameter were subjected to single cell sorting into the wells of 96-well PCR plates. Whole genome amplification was then carried out on the cells, followed by amplification verification via PCR for the KLC2 gene. This verification step allowed us to limit the whole exome sequencing to only those wells that contained amplified DNA. The number of successfully amplified cells per sample has ranged from 20 to 70. After amplification, the DNA was transferred for sequencing and the data were returned to us for analysis.

The amplification and sequencing subtasks are considered to be complete unless resequencing needs to be done, and/or if the phylogenetic tree analysis indicates we need to analyze more cells/patient or more patients. The sequence analysis has been started and is ongoing. The sequence analysis performed to date is described below:

Single cell sequence analysis

We obtained single cell exome sequence data from 384 samples. We used PICARD (Broad Institute, Cambridge MA), a set of command line tools for manipulating high-throughput sequencing data, to gather on-target base quantity and depth for quality control. Samples with on-target base quantity lower than $1.1x10^9$ were removed from further analysis (**Fig. *1***). The remaining samples on have average 49% bases with depth larger or equal to 10.

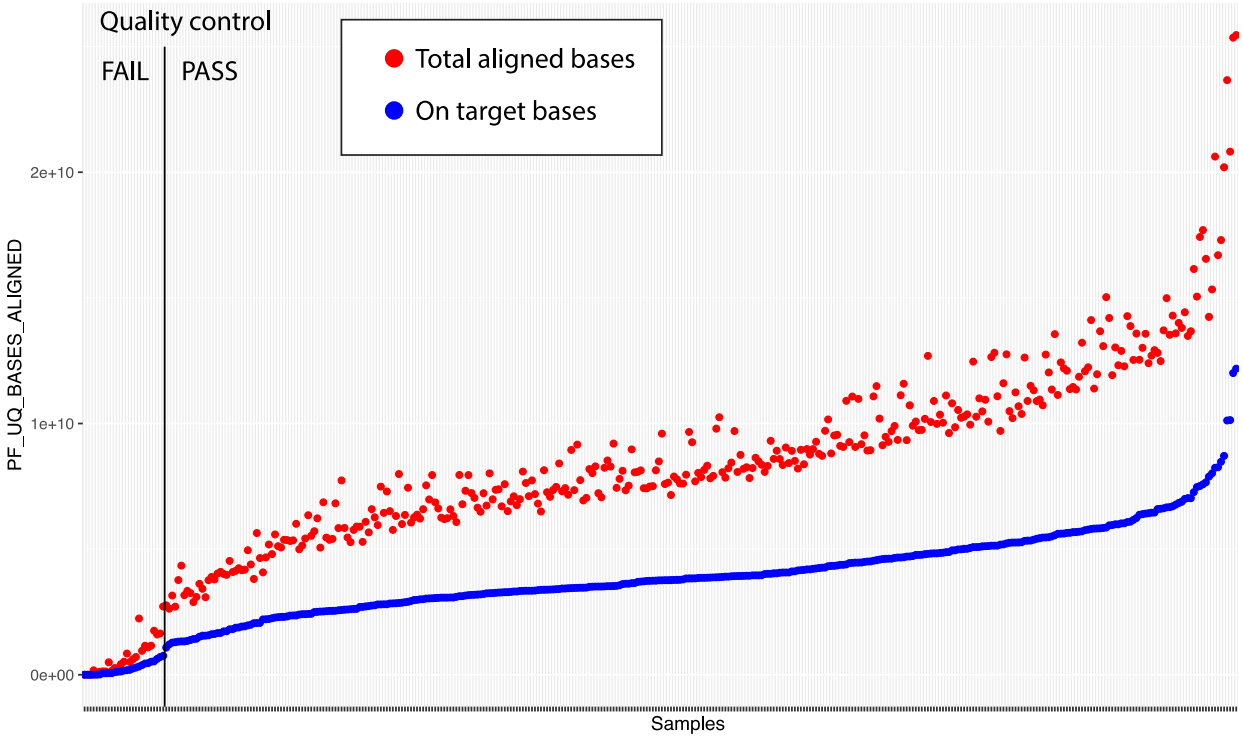

Fig. 1 Total and on-target bases of samples.

We then followed Genome Analysis Toolkit (GATK) best practice (1) to perform variant calling. We used default filters and performed joint calling for each patient and for all samples together. We have finished initial variant calling on 279 samples from 10 patients so far. The number of samples and variants are shown in **Table 1**.

To validate the identity of each sample (i.e., single cell sequence) and avoid label mixing, we used hierarchical clustering for 279 good quality sequencing samples with 376 select variant sites. These sites were selected based on: 1. Passing all default filters; 2. Coverage of the site by >20 samples; 3. Existence of only one alternative allele. The clustering and heatmap shows samples from the same patient cluster together in general with a few exceptions (
Fig. *2*). We plan to further validate their identities with germline mutations from normal control samples.

Table 1 Number of samples and variants

Patient	#Sample	#Total_variants_PASS
17004	70	2498656
17028	59	2343082

17029	42	1792602
17017	36	1268016
17008	23	1409957
16011	18	1087513
17005	14	865748
17011	11	673366
18001	9	775233
17030	3	291020

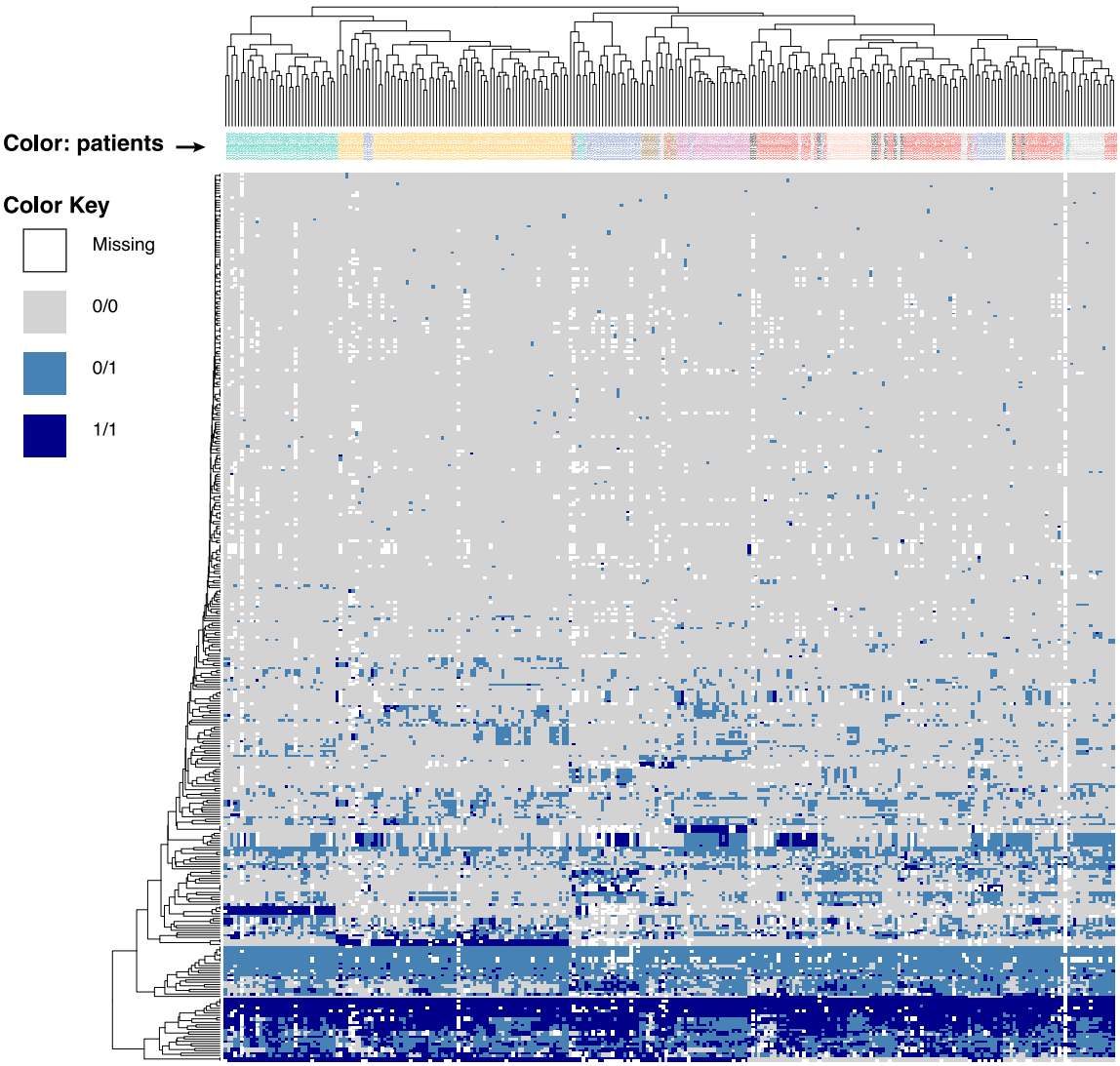

Fig. 2 Heatmap of 376 sites for all samples

Specific Aim III. Using the whole exome sequence data, analyze phylogenetic relationships of tumor cells and develop a prognostic classifier.

Major Task 4. Reconstruct phylogenetic trees.
Major Task 5. Compare phylogenetic trees of groups with and without metastasis and develop a prognostic classifier.
We are currently at the start of **Major Task 4**. We built a phylogeny for one of the patients as an example to test our tree-building pipeline. We first identified a high confidence set of variants with filters 1-3 above, and

added two more filters: 4. Single nucleotide variation only, no insertion/deletion; and 5. Genotype is either 0/0 or 0/1.

After filtering we identified 2569 sites and we built a maximum parsimony tree with mid-point rooting. We used 100 bootstraps to assess the node support. The result (**Fig. 3**) shows a decently-supported balanced tree with long tips branch lengths and short internal branch lengths, which may suggest an expanding tumor cell population.

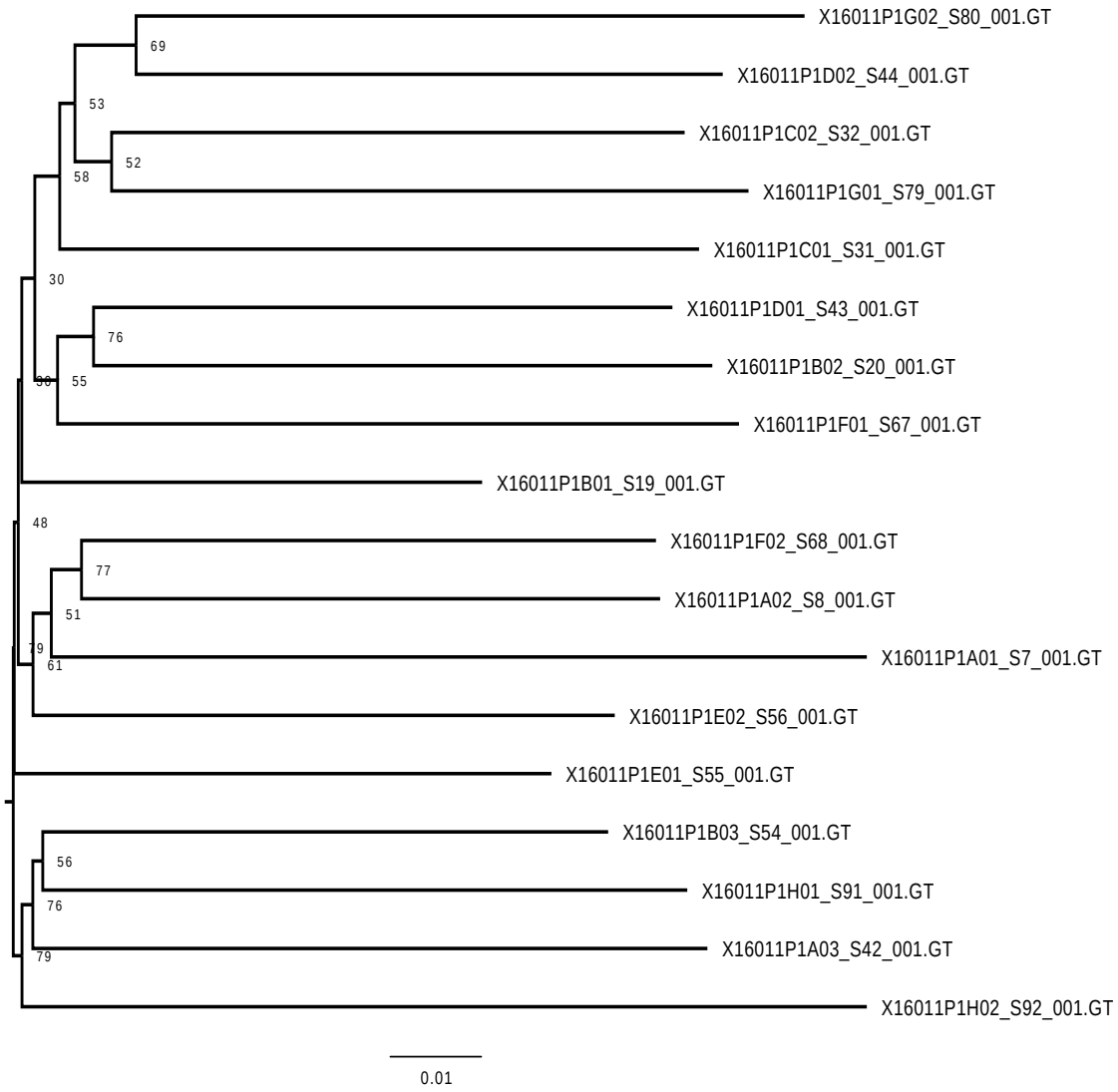

Fig. 3 Phylogeny for patient 16011. The scale bar is in units of nucleotide substitutions per site.

Our next step is to further improve our variant calling and tree building pipeline to achieve more stable trees with stronger support. We will then build phylogenies and estimate population genetic parameters for each patient. Finally, we will test if the tree dynamics and parameters are correlated with patient clinical outcomes. Now that DNA sequencing of the training set has been completed, we anticipate completion of the project within the next eight months.

Specific Aim IV. Validate the prognostic classifier developed in Specific Aim III in an independent blinded study.
Nothing to report at this time.

Reference

1. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.

Opportunities for training and professional development
Dr. Yuantong Ding developed computational models for the development of phylogenetic trees for her PhD thesis with Dr. Allen Rodrigo, in part using data from this project. Her thesis title is: Application of Phylogenetic Analysis in Cancer Evolution (Duke, University, 2018). She is now a Postdoctoral Fellow and is continuing to use the models she developed in order to complete this project.

How results were disseminated to communities of interest
Nothing to Report.

Plans for next reporting period
In the next reporting period, we plan to develop a prognostic classifier, completing Specific Aim III. We will then begin Specific Aim IV to validate the prognostic classifier.

4. Impact

Impact of the development of the principal discipline of the project
Nothing to Report

Impact on other disciplines
Nothing to Report

Impact on technology transfer
Nothing to Report

Impact on society beyond science and technology
Nothing to Report

5. Changes/Problems
Nothing to Report

6. Products
None to date

7. Participants & Other Collaborating Organizations
E.B. Gottlin, investigator, performed tumor cell isolation, 2.4 cal. months
S.G. Gregory, investigator, supervised WGA, 0.48 cal. months
E.F. Patz, Jr., PI, 1.4 cal. months
E.A. Burns, Lab Assistant, 2.4 cal. months
A.G. Rodrigo, 0.15 cal. months
Y. Ding, Postdoctoral Associate, 0.15 cal. months

LabCorp is doing the DNA sequencing for this project; there has been no change in the active support of the PI.

8. Special Reporting Requirements
None

9. Appendices
None