AWARD NUMBER: W81XWH-15-1-0395

TITLE: Integrated Genomic Biomarkers to Identify Aggressive Disease in African Americans with Prostate Cancer

PRINCIPAL INVESTIGATOR: Dr. Albert Levin

CONTRACTING ORGANIZATION: Henry Ford Health System

REPORT DATE: Nov 2019

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Development 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
Nov 2019	Final	01 Sep 2015 – 31 Aug 2019

4. TITLE AND SUBTITLE
Integrated Genomic Biomarkers to Identify Aggressive Disease

In African Americans with Prostate Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-15-1-0395

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Albert Levin

E-Mail: alevin1@hfhs.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Henry Ford Health System
1 Ford Place
Detroit, MI 48202-3450

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development 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

Robust methods for risk stratification of prostate tumors are needed to enable men and their physicians to safely select between post-treatment surveillance and immediate adjuvant therapy. The purpose of our research is to use a mulit-omic approach to identify somatic copy number alterations and methylation markers in the primary tumors of African American men that can serve as a component of their recurrence risk assessment and be applied in treatment planning to help reduce the racially disparate rates of mortality from CaP. Through whole genome copy number alteration and methylation scans, the study will identify individual and integrated DNA-based biomarkers of biochemical recurrence in 200 African American men (100 with and 100 without biochemical recurrence). These biomarkers will then be validated in an independent set of 200 African American men.

15. SUBJECT TERMS
Prostate cancer; DNA; copy number alterations; methylation; biomarker; racial disparities; integrative.

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	22	19b. TELEPHONE NUMBER *(include area code)*

TABLE OF CONTENTS

1. Introduction

Approximately ~33,000 men die each year from prostate cancer (CaP), in particular from disease recurrence. African American men have higher CaP mortality rates than age matched European American males. Risk of disease recurrence after primary treatment is difficult to predict with clinical variables and prostate specific antigen. Robust methods for risk stratification of prostate tumors are needed to enable men and their physicians to safely select between post-treatment surveillance and immediate adjuvant therapy. The purpose of our research is to use a mulit-omic approach to identify somatic copy number alterations and methylation markers in the primary tumors of African American men that can serve as a component of their recurrence risk assessment and be applied in treatment planning to help reduce the racially disparate rates of mortality from CaP. Through whole genome copy number alteration and methylation scans, the study will identify individual and integrated DNA-based biomarkers of biochemical recurrence in 200 African American men (100 with and 100 without biochemical recurrence). These biomarkers will then be validated in an independent set of 200 African American men.

2. Keywords

prostate cancer; DNA; copy number alterations; methylation; biomarker; racial disparities; integrative.

3. Accomplishments

Major Task #1: Identify subjects and tissue specimens for biomarker discovery and validation.

- Identify from the existing database at Henry Ford Health System (HFHS) lists of eligible prostatectomy patients as defined in the Research Strategy, confirm availability of banked formalin fixed and paraffin-embedded (FFPE) prostate tissue with biorepositories. Target completion January 31st 2016; Completed January 1st 2016

- Calculate CAPRA-S scores for all eligible subjects. Target completion January 31st 2016; Completed January 15th 2016

- Perform incidence sampling to determine discovery and validation study samples. Target completion September 1st 2016; Discovery sample 100% completed September 1st 2017

- Tumor blocks will be pulled from archive, determination of the optimal block, and sections cut and tumor areas marked by pathologist. Target completion September 1st 2017; Discovery sample pathologic review completed September 1st 2017; Validation sample pathologic review completed February 10th 2019

- Pathology review completed of cut cases and slides transferred to UCSF. Target completion January 1st 2018; Discovery sample 100% completed as of September 1st 2018. Validation sample 100% completed as of February 18th 2019

Major Task #2: Tissue processing and DNA extraction for entire project.

- Manual tumor tissue macrodissection. Target completion January 1st 2018; Discovery sample 100% completed as of January 15th 2019. Validation sample 100% completed as of March 15th 2019.

- DNA extraction and quality assessment. Target completion March 31st 2018; Discovery sample 100% completed as of Febuary 1st 2019; Validation sample completed as of May 15th 2019.

Major Task #3: Perform genomic microarray experiments.

- Carry out array comparative genomic hybridization (aCGH) on Aim 1 DNAs at UCSF. Target completion date September 1st 2017; Agilent reagent quotes obtained. Reagents ordered. 60 samples run and passed QC as of September 1st 2017. As of March 1st 2019 and after thorough comparative analysis between the copy number derived from aCGH and the Illumina EPIC arrays, we have determined that the quality of the copy number alteration calls from the methylation arrays was of unacceptably low sensitivity and specificity to justify their use.

- Quality control for aCGH and determination of copy number via CBS. Target completion date September 1st 2017; 60 samples run and passed QC as of September 1st 2017. We are currently running the aCGH arrays on the remainder of the discovery specimens. This process began in August 2019, and our current target completion date of February 28th 2020.

- Quality control of methylation microarray data and preparation of an analysis dataset. Target completion date September 1st 2017; Our full sample of 200 discovery subjects was completed and return on July 1st 2019, Quality check conducted and initial analyses have been completed (see next section) as of July 2019.

- Conduct methylation microarray experiments on Aim 1 DNAs. Target completion date September 1st 2017; Worked out an agreement with Illumina to provide methylation reagents for 1st 48 samples to determine if the Illumina EPIC arrays are going to be useful in also identifying copy number alterations in FFPE preserved prostate tumors from African Americans. The initial results of those analyses did not prove promising, and as part of our second wave of arrays, we included a set of 8 African American prostate normal reference specimens which we believe would improve the quality of the copy number calls from the methylation arrays. Illumina has again provided the additional arrays and reagents for the normal specimens to determine the degree to which these will help facilitate a higher fidelity of copy number calling from methylation arrays. Unfortunately, the results of this endeavor has revealed sub-optimal performance of the EPIC arrays for calling copy number alterations, and we have returned to typing copy number alterations using the aCGH arrays.

Major Task #4: Statistical analyses for GEMCaP and published methylation biomarkers. Target completion November 1st 2017; Based on our initial aCGH discovery sample set, preliminary analyses for GEMCAP as of September 21st 2017. Based on our finalized methylation array discovery sample, we have completed our analysis of published methylation biomarkers of biochemical recurrence as of August 1st 2019

Major Task #5: Discovery of African American specific copy number and methylation biomarkers. Target completion Febuary 1st 2018. Discovery and cross-validation of methylation biomarker completed on September 1st 2019

- We have used TCGA data to identify a copy number based biomarker associated with recurrent disease. The manuscript from these analyses has been re-submitted to BMC Genomics as of November 1st 2019.

- Using the initial sample of aCGH data, we have conducted a preliminary genome-wide discovery analysis of copy number alterations associated with biochemical recurrence in African Americans. Analysis completed as of September 1st 2019.

- Using our African American discovery sample, we have identified and cross-validated a multi-CpG site methylation predictor of biochemical recurrence. There are 19 CpG sites in this biomarker, and the cross-validated concordance index (i.e. prediction accuracy) is 73%.

- Using the African American prostate cancer TCGA data, we have used the methylation and RNA-Seq data to determine the genes whose expression is influenced by alterations in these 19 CpGs. Further, we have used this data to a) characterize CpG to gene expression associations across the entire genome and b) determined how these relationships differ between the somatic genomes of African Americans and European Americans.

- Using the African American prostate cancer TCGA data, we have used the methylation and RNA-Seq data to construct methylation based models of gene expression in both African Americans and European Americans, and we are currently using those models to a) predict gene expression in our discovery sample of African Americans with methylation data and b) to use those predicted expression values to construct a more powerful gene-based statistical test for methylome wide association studies.

Major Task #6: Validate integrated biomarker panel in a separate discovery set of African American prostate cancer. Target completion July 1st 2018; not yet started.

- While we are still waiting for the final set of aCGH data to complete, we have conducted preliminary analysis of histopathologic features digitally extract from high powered magnification (40X) images of the H&E slides from our African American cohort. The intent of these analyses was to determine whether such features in aggregate could further augment the predictive ability of the methylation biomarker discovered in Major Task #5. Our preliminary findings not only show that these features can augment the multi-CpG site biomarker of biochemical recurrence for prediction of early recurrence (before 3 years) but also that there combination renders Gleason Grade and CAPRA-S scores insignificant in the models, suggesting that they are better predictors of early recurrence in African Americans than these established clinicopathologic predictors. These preliminary results form the core of an IDEA expansion grant submitted to the CDMRP in September 2019.

Major Task #7: Draft manuscripts for publication.

Manuscript #1: Performance of existing clinical pathologic predictors of prostate cancer in an African American population. 90% completed as of September 1st 2019.

Manuscript #2: Prostate and breast cancers harbor common somatic copy number alterations that consistently differ by race-ethnicity. Target completion November 2017; 100% completed as of September 28th 2018. This paper was submitted to BMC Genomic on November 1st 2019.

Manuscript #3: Identification of methylation-based predictors of aggressive prostate cancer in African American men. 50% completed as of November 1st 2019.

Manuscript #4: African American prostate cancer recurrence risk prediction using both genome-wide somatic methylation and histopathologic image features. 30% completed as of November 1st 2019.

Manuscript #5: Differential effect of methylation on gene expression in African American and European American prostate tumors. 40% completed as of November 1st 2019.

Manuscript #6: Gene-based epigenome-wide association mapping of cancer outcomes using somatic reference RNA-Seq data. 20% completed as of November 1st 2019

What was accomplished under these goals?

For our sub-cohort of 200 AA men with FFPE prostate tissue and clinical data (CAPRA-S components and BCR status) from our parent DoD award, DNA was extracted from the macro-dissected primary tumor focus for each patient, and genome-wide methylation was examined using the Epic array (Illumina), which consists of ~850k methylation CpG sites. Sample-wise quality control removed 18 subject from the final analytic sample of 182 subject, and CpG site quality control reduced the number of analyzable CpG sites to 658,115. The clinical/pathologic characteristics of this sub-cohort relative to the complete cohort are presented in Table 1.

Table 1: Demographic characteristics of the Henry Ford Health System African American prostatectomy overall cohort and the study cohort (1999-2012)

	Overall cohort			Study Cohort w/ EPIC CpG Data		
Characteristic	**No BCR (n=592)**	**BCR (n=282)**	**p-value**	**No BCR (n=73)**	**BCR (n=129)**	**p-value**
Mean Age at Diagnosis	60.45±0.33	60.73±0.45	0.214	60.72 ± 0.89	60.67 ± 0.60	
Mean Age at Prostatectomy	60.85±0.32	61.05±0.44	0.244	61.40 ± 0.843	60.99 ± 0.60	
PSA at Diagnosis	6.45±0.20	9.46±0.47	<0.001	6.91 ± 0.63	9.59 ± 0.68	0.005
Grade Group			<0.001[a]			0.002[a]
1	173 (29.2%)	45 (16.0%)		21 (29.6%)	18 (14.0%)	
2	269 (45.4%)	87 (30.9%)		27 (38.0%)	40 (31.0%)	
3	91 (15.4%)	63 (22.3%)		9 (12.7%)	32 (24.8%)	
4	58 (9.8%)	86 (30.5%)		14 (19.7%)	39 (30.2%)	
Unknown	1 (0.2%)	1 (0.3%)		0 (0.0%)	0 (0.0%)	
Pathological Stage			<0.001[b]			0.002[b]
2A	57 (9.6%)	15 (5.3%)		6 (8.5%)	5 (3.9%)	
2B	122 (20.6%)	95 (33.7%)		25 (35.2%)	46 (35.7%)	
2C	267 (45.1%)	42 (14.9%)		21 (29.6%)	14 (10.9%)	
3A	118 (19.9%)	70 (24.8%)		16 (22.5%)	36 (27.9%)	
3B	23 (3.9%)	54 (19.1%)		2 (2.8%)	24 (18.6%)	
4	1 (0.2%)	6 (2.1%)		1 (1.4%)	4 (3.1%)	
Unknown	4 (0.7%)	0 (0.0%)		0 (0.0%)	0 (0.0%)	

Abbreviations: BCR, denotes biochemical recurrence; PSA, prostate specific antigen.

[a] Grade group 1-2 and 3-4 combined for Chi-Square test

[b] Pathological T Stage 2A-2C and 3A-4 combined for Chi-Square test

CpG site Beta-value profiles for each subject underwent quality control and normalization using the "minfi" pipeline. The resulting Beta-values were transformed to M-values [log(Beta/(1-Beta)] prior to analysis. Association between single CpG sites and biochemical recurrence (BCR) was performed using Cox proportional hazards regression.

Our initial Aim 1 analyses were restricted to the validation of 25 loci previously reported where hypermethylation was associated with increased risk of BCR in European American men (Mahapatra et al. 2012). In our data, all of the hazard ratios (HR) reflected increased risk with increasing methylation (i.e. HRs > 1), and the range of the HRs was from 1.11 to 1.42, per unit increase in the M-value (Table 2).

Table 2: African American evaluation of CpG sites identified by Mahapatra et al. where hypermethylation was associated with biochemical recurrence.

Illumina ID	CHR	BP	GENE	HR	P-value
cg06377278	1	25256369	RUNX3	1.24	0.001
cg15457058	1	47882322	FOXE3	1.29	$5.03*10^{-5}$
cg11914824	1	203320297	FMOD	1.29	0.004
cg14174946	2	31805218	SRD5A2	1.29	0.002
cg17108819	2	87017953	CD8A	1.31	$4.72*10^{-5}$
cg19443257	3	140770549	SPSB4	1.41	$3.91*10^{-5}$
cg08894629	5	523529	SLC9A3	1.32	0.004
cg02625102	5	80256351	RASGRF2	1.30	$5.21*10^{-5}$
cg17772342	5	134871686	NEUROG1	1.42	$2.30*10^{-4}$
cg14249876	7	44143993	AEBP1	1.24	$3.82*10^{-4}$
cg26990587	7	100253901	ACTL6B	1.19	0.007
cg02661879	7	128470614	FLNC	1.11	0.012
cg00234261	7	151106082	WDR86	1.33	$5.87*10^{-4}$
cg24914278	8	109799739	TMEM74	1.24	0.004
cg22154482	9	95820922	SUSD3	1.19	0.081
cg02308884	10	71892293	AIFM2	1.26	0.001
cg13467162	11	65600888	SNX32	1.27	0.002
cg24068708	11	75236459	GDPD5	1.13	0.116
cg18232816	12	33591505	SYT10	1.15	0.110
cg25823578	14	52781312	PTGER2	1.21	0.017
cg06785999	14	60975964	SIX6	1.27	$4.05*10^{-6}$
cg13641815	14	105953421	CRIP1	1.15	0.059
cg18422586	16	22825856	HS3ST2	1.19	0.010
cg13991233	17	10102558	GAS7	1.18	0.086
cg13140564	20	44540827	PLTP	1.17	0.001

Abbreviations: HR, denotes hazard ratio; CHR, chromosome; BP, base pair position in human genome build 19.

Of the 25 loci, 20 (80%) and 17 (68%) were validated at $p<0.05$ and $p<0.01$ thresholds, respectively. The most significant result was identified at the SIX6 gene locus (HR=1.27; $p=4.05*10^{-6}$). These findings are remarkably consistent with the original paper findings and reveal the applicability of these CpG biomarkers, discovered in EAs, to AAs.

In addition to the validation of methylation biomarkers of aggressive CaP from the existing literature, we have conducted biomarker discovery in our African American sub-cohort (Aim 2). To build a BCR prediction model, we used Cox proportional hazards regression with both Lasso and Elastic Net penalized regularization[40] to remove CpGs that there were not highly associated with BCR. To overcome the problem of an initial large number of CpGs (n=658,115 post quality control), we used the Bhattacharyya Distance (BD), mutual information, and t-tests to perform initial CpG feature screening. BD is a univariate filter methods that ignores feature dependencies and autocorrelation and uses the mean and variance of the class conditional distributions to rank features. After performing feature selection, the top 2,000 CpG sites were passed to the penalized Cox model for further feature selection and model building using two-layers of cross-validation. We used an inner layer of leave-one-out cross-validation to estimate the optimal shrinkage parameter for the model and an outer layer to calculate a CpG score $\sum_{j=1}^{n} \beta_j * CpG_{i,j}$, where $CpG_{i,j}$ is the M-value for the *j*-th CpG site for the *i*-th individual left out of each round of the model building process. We used these CpG scores to generate an unbiased estimate of Harrell's concordance index, which is a measure of prediction accuracy for survival models. Our best model of the six evaluated (3 features selection procedures multiplied by the 2 penalized regression models) was the BD features selection combined with the Lasso penalization. This final model consisted of 19 CpGs, 13 of which were located within the following genes: ETV1, PAX5, TESC, PTPRN2, CBFA2T3, MIAT, LRRC8D, ZNF606, THAP7, LEPR, MUC6, MAN1B1, and PAX3. This model had a cross-validated concordance index of 72%, and the univariate and multi-variate effects for these CpGs are presented in Table 3.

Table 3: Hazard ratios for the CpG sites selected for inclusion in the multi-CpG site model associated with risk of biochemical recurrence risk in African Americans.

		Single CpG Site Models			**Multi-CpG Site Model**	
Illumina ID	**Gene Distance/Expression***	**HR**	**SE**	**P-value**	**HR**	**SE**
cg13910860	PAX5/CLTA (15)	1.37	0.08	2.61E-05	1.21	0.10
cg12801791	TESC/C12orf49 (12)	0.72	0.14	0.017314	0.60	0.16
cg25198113	-/ESPN (16)	0.77	0.15	0.072689	0.77	0.16
cg10218605	PTPRN2/-	1.36	0.08	0.000274	0.99	0.09
cg27130240	CBFA2T3/ACSF3 (25)	0.67	0.11	0.000339	0.91	0.15
cg03612722	MIAT/-	1.38	0.08	5.48E-05	1.09	0.10
cg16422731	-/-	0.61	0.13	0.00018	0.71	0.16

cg04168853	-/ACTN4(27)	1.39	0.08	1.93E-05	1.27	0.09
cg02737895	LRRC8D/-	1.55	0.13	0.000694	1.35	0.14
cg12689060	LOC100128398	1.32	0.08	0.000426	1.09	0.08
cg04802407	THAP7/SLC7A4 (21)	1.56	0.12	0.000217	1.39	0.14
cg08234308	LEPR/-	1.20	0.07	0.005346	1.01	0.07
cg00029701	-/-	0.58	0.12	1.35E-05	0.68	0.14
cg10047905	-/-	0.70	0.10	0.000314	0.97	0.13
cg13892862	MUC6/B4GALNT4 (44)	0.14	0.62	0.001575	0.75	0.64
cg20281632	ETV1/-	1.46	0.16	0.01949	1.09	0.19
cg09325123	MAN1B1/-	0.57	0.14	7.59E-05	0.66	0.16
cg02347074	PAX3/FARSB (9)	1.42	0.14	0.010867	1.17	0.17
cg11791812	-/-	0.42	0.23	0.000161	1.10	0.23

Abbreviations: HR, denotes hazard ratio; SE, standard error.
*Gene assignment based on proximity to the CpG site ("Distance) and association with gene expression ("Expression") based on the African American prostate cancer cases from TCGA. For the "Expression" assignment, the most significant single gene is listed in the Table, and in parentheses, the total number of genes with expression significantly affected (after conservative Bonferroni adjustment for 8 million tests) is listed. A "-" mark indicates that no gene could be assigned.

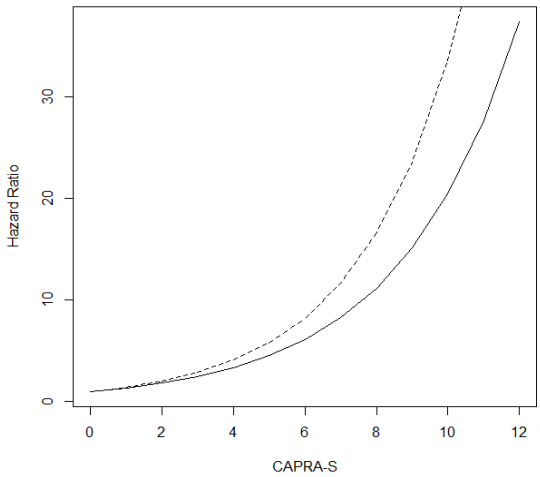
The gene assignments for each of the CpG sites that appear in Tables 2 and 3 are all based on the base-pair position of the CpG sites. In other words, the most proximal gene is assigned to each of the CpG sites. However, biologically, it is plausible that methylation levels at a CpG site could affect the expression of multiple genes, some of which may be at a distance from that CpG site. In order to biologically assign genes to CpG sites, we again used TCGA data to determine which genes have expression values that are significantly correlated with each of the CpG sites. For each CpG site, we tested correlation with gene expression for all genes within 1Mb of each site. Given that TCGA has methylation data on the Illumina 450k array, this assessment could only be done for a subset of the CpG sites from Table 3. Specifically, there were 8 total CpGs where expression assignment could be conducted (i.e. on both the 450k and EPIC arrays). After strict Bonferroni correction for nearly 8 million tests across the entire genome (all genes within 1Mb of each CpG site on the 450k array), these 8 CpG sites were significantly associated with the expression of 21 genes on average (minimum=9 genes and maximum=44 genes). For one CpG (cg04168853), a gene assignment could only be made by the expression. For the remaining seven, the most significantly associated gene based on expression was not the most proximal gene based on distance. These findings have consequences for how we follow up these CpG sites biologically. In other words, the effect of CpG sites is likely to be primarily through gene expression, and efforts to mechanistically understand their effects on recurrence risk in African American men would be miss-guided or limited based on distance based gene assignment only or expression assignment using data from European American men (see below). However, through examining both the African American and

European American associations between methylation sites and gene expression, it became apparent that our results had important additional consequences. First, there were striking differences in these association by race in the prostate cancer TCGA dataset. Specifically, after Bonferroni adjustment for ~8 million tests ($p<6.25*10^{-9}$), 4,270,064 and 5,037,796 genes were significantly associated with CpG sites in African American and European American TCGA prostate tumors, respectively. Moreover, 4,034,397 had significantly different associations in African Americans and European Americans. These findings have not been previously appreciated in the literature and underscore further genetic differences between African American and European American prostate tumors and the need for larger genomic studies of African American prostate tumors to be carried out.

Second, in assessing the association between methylation and gene expression, we realized that these results may have broader consequence for epigenome-wide studies of cancer outcomes in general. Similar to what has been done with germline genetic data using resources such as GTEx and methods such as PrediXcan, it is possible to construct models to predict gene expression based on methylation in tumors. Those models can be used for the prediction of gene expression in datasets that have genome-wide methylation data only, and these predictions can be used as gene-based tests of epigenetic associations. In doing so, such gene-based tests could reduce the multiple testing burden by an order of magnitude and increase power. We have used penalized Poisson regression models and cross-validation to construct and evaluate these models in African American and European American TCGA prostate tumors. Our cross-validated results suggest that these models capture a large portion of the variability in gene expression in both African Americans (median pseudo R^2=0.46, 25^{th}-75^{th} percentile 0.08-0.71) and European Americans (median pseudo R^2=0.44, 25^{th}-75^{th} percentile 0.27-0.56). For our project, our next step is to deploy these models in our own data and use these predicted expression values to construct new models of BCR prediction and compare the concordance indices with our single site model (Table 3). Further, we are currently writing up the findings for both of these analyses for submission, expanding the scope beyond prostate cancer.

Using data from the full HFHS radical prostatectomy cohort of both African American (n=874) and European men (n=1,128) from 1999-2012, we evaluated the comparative BCR predictive ability of Cancer of the Prostate Risk Assessment (CAPRA-S) tool in these two race-ethnicities. Across the two ethnicities, African American men tended to have higher CAPRA-S scores relative to European American men ($p=5.71*10^{-4}$). Despite this disparity of more aggressive disease, African Americans had a predicted lower risk of recurrence at each CAPRA-S score relative to European Americans (p=0.061; Figure 1), with the largest differences evident at the highest CAPRA-S risk levels (CAPRA-S > 9; p=0.021). Consistently, the concordance index for the CAPRA-S model was higher for European American (77%, 95%CI 74%-81%) in comparison to African American men (73%, 95%CI 69%-76%). These findings collectively demonstrate that existing clinicopathologic parameters may be suboptimal for risk stratification for African American men and that more precise prediction of aggressive disease in African Americans is needed.

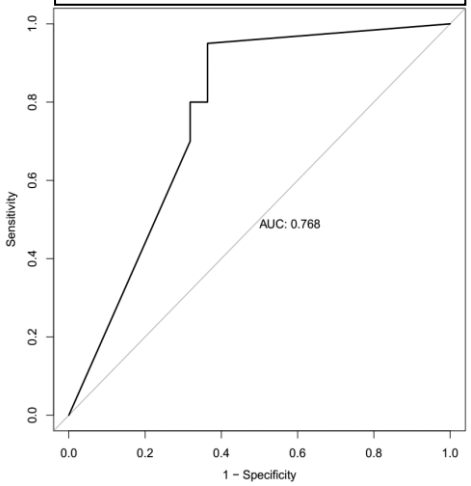
Figure 1: Hazard ratios at each CAPRA-S score for European American (dashed line) and African American (solid line) men from HFHS.



With the intent to improve the accuracy of recurrence prediction in African American men, we also evaluated the predictive ability of histopathologic image features extracted from the primary tumor to predict early BCR. Tumor tissue slides were digitalized using Aperio CS2 digital pathology slide scanner at 40x magnification for 20 patients with early BCR (BCR <3 years of surgery) and 22 patients without early BCR (no BCR <5 years of surgery). For each tumor, five regions were sampled at random within the pathologist marked primary tumor focus. The cell morphological features (cell, cytoplasm, and nuclei) for those regions were extracted using CellProfiler. Features were aggregated across regions by mean, median, standard deviation and deciles (10-quantiles) of extracted values, resulting to 14,380 aggregated features. The top 1,000 features were

selected using the information gain ratio as calculated within FSelector, and those features were used to construct a Random Forest (RF) classifier. Leave-one-out cross-validation was used to calculate an unbiased estimate of the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC). Specifically, in one run, the features from all five regions from one patient were reserved as testing while the remainder of the data was used to train the RF model. The trained model was used make binary predictions for each of the five regions for the left-out tumor, and these predictions were summarized into a recurrence probability (proportion of predicted early recurrence regions). This process was repeated so that all 42 patients will have one recurrence probability. Figure 1 displays the resulting cross-validated ROC curve, which had an AUC of 0.77. By comparison, the AUC for a model based on CAPRA-S (which includes Gleason grade and other existing clinicopathologic predictors of BCR) was 0.71 for this dataset, which demonstrates that there is more recurrent disease information to be gained from existing histopathologic images.

Figure 2: ROC for prediction of early BCR based on digitally extracted tumor histologic imaging features

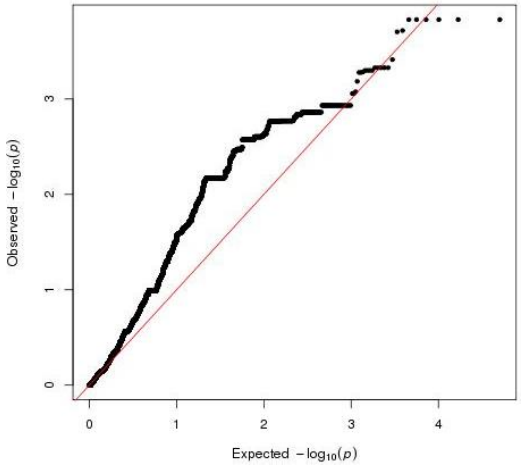


Finally, in an integrated model including both the CpG score from our newly discovered biomarker (Table 3) and histopathologic image feature score ("Histo Score"), increasing levels of the CpG (p=0.018) and image feature (p=0.016) scores were both significantly associated with increased risk of early BCR (Table 4; Model 1). Moreover, the AUC for this model was 89%. Inclusion of the CAPRA-S score increased the AUC to 91%, and consistently, CAPRA-S was not a significant predictor (p=0.112; Table 4, Model 3) after inclusion of the CpG and image feature scores. Clearly, methylation data and histopathologic features improve upon clinicopathologic parameters such as CAPRA-S in predicting BCR in African American prostate cancer. Given these encouraging preliminary findings, our plan is to expand this sample size to include all 182 subjects with high quality methylation data.

Table 4: Integrated models of CpG and histopathologic image feature scores are associated with BCR in African American men independent of known clinical and pathologic predictors.

	Model 1				**Model 2**				**Model 3**		
	OR	95% CI	P		OR	95% CI	P		OR	95% CI	P
Histo Score	13.54	1.63-112.72	0.016		13.37	1.59-112.61	0.017		9.06	0.99-83.27	0.051
CpG Score	7.55	1.42-40.06	0.018		7.41	1.36-40.56	0.021		4.70	0.85-25.97	0.076
Gleason Grade	-	-	-		1.05	0.47-2.31	0.910		-	-	-
CAPRA-S	-	-	-		-	-	-		1.39	0.92-2.10	0.115

Figure 3: QQ plot of individual gene p-values for the association between copy number and biochemical recurrence in the initial sample of African American men with aCGH data.



While our attempt to use the copy number alteration calls from the Illumina EPIC array was not successful (median gene specific sensitivity and specificity < 0.6), we have moved forward with analyses of association between DNA copy number alterations at individual genes (n=25,247) and biochemical recurrence based on our initial sample and in advance of receiving our final array comparative genomic hybridization data for the full discovery sample. The quantile-quantile plot of the observed and expected p-values from this analysis is presented in Figure 3. There were a total of 1,297 genes that were significant at a p-value threshold of 0.01, and all of these genes had a false discovery rate < 20%. Among these were 12 genes that were part of the GEMCaP biomarker. All 12 of these genes had a consistent direction of effect with the original biomarker (3 and 9 where amplification and deletion were respectively associated with increased risk of recurrence). These findings suggest that there are true copy number alterations that are associated with prostate recurrence in this African American data set.

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

Dr. Paris at UCSF had a summer intern who was part of the UCSF Minority Training Program in Cancer Research. This DoD project allowed the student to gain experience in pathology review, macrodissection and DNA extraction. She also participated in monthly UCSF-Henry Ford team meetings.

How were the results disseminated to communities of interest?

We have recently started to report our findings to the scientific community. In early November, Dr. Paris and her intern present our methylation discovery and literature validation findings from our study at the UCSF Urology Research Symposium, and Dr. Levin presented these findings along with the integrated histopathologic image features work at an invited Cancer Center Biostatistics seminar within the Biostatistics Department at the University of Michigan. Our race-differentiated copy number alteration paper is currently under review at BMC Genomics.

What do you plan to do during the next reporting period to accomplish the goals?

Nothing to report

4. IMPACT

a. **What was the impact on the development of the principal discipline(s) of the project?**
Our initial findings from genome-wide copy number and methylation data suggest that some molecular biomarkers of biochemical recurrence discovered in European American will apply to African American men. However, differences are also apparent and justify discovery of ethnic specific markers African American men, which we now have established through the discovery and cross-validation of an African American methylation-based biomarker of recurrence risk.

Through the use of TCGA data to assign CpGs to genes via gene expression associations, we identified a large number of highly significant difference in the association between degree of CpG methylation and gene expression by race in prostate cancer. These significant interactions again emphasize the need for additional and

larger genomic studies of African American prostate cancer. These findings further emphasize that the gene-based epigenome-wide analysis approach that we have developed needs to be applied in a race specific manner.

b. **What was the impact on other disciplines?**

Our manuscript using TCGA data that details the race-differentiated copy number alterations that are unique and shared between prostate and breast has a clear impact on the field of breast cancer racial disparities. In this same manuscript, we developed a new area under the curve method for quantifying copy number alterations and testing with outcomes. This new approach could be used in the analysis of copy number alterations in any tumor type and therefore has impact on cancer research in general.

The methylation site associations with gene expression and the methylation-based gene expression prediction models will likely have an impact on other cancer disciplines. First, our results for race-dependent effects of methylation on gene expression have not yet been reported and exist for prostate and breast tumors. Second, the predictive gene expression models will provide a completely new method for performing epigenome wide association analyses of cancer outcomes for any cancer, not just prostate.

What was the impact on technology transfer?
Nothing to report

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

Nothing to report

5. **CHANGES/PROBLEMS:**
a. **Changes in approach and reasons for change**

Nothing to report

b. **Actual or anticipated problems or delays and actions or plans to resolve them**

Nothing to report

c. **Changes that had a significant impact on expenditures**

We added two staff members (Dr. Indra Adrianto and Ms. Cara Canella) to aid us in the analysis of data for this study.

d. **Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents**

Nothing to report

e. **Significant changes in use or care of human subjects**

Nothing to report
f. **Significant changes in use or care of vertebrate animals.**

Nothing to report

g. **Significant changes in use of biohazards and/or select agents**

Nothing to report

6. **PRODUCTS:**

a. **Publications, conference papers, and presentations**

Presentation #1 DoD IMPaCT Meeting 2016, Bethesda, MD: The impact of self-identified race-ethnicity and genetic ancestry on a commonly used clinicopathologic predictor of biochemically recurrent prostate cancer.

Presentation #2 AACR Meeting 2018, Chicago, IL: Breast and prostate cancers harbor common somatic copy number alterations that consistently differ by race.

Presentation #3 AACR Disparities Meeting 2019, San Francisco, CA: Breast and prostate cancers harbor common somatic copy number alterations that consistently differ by race and are associated with survival.

Presentation #4 UCSF Urology Research Meeting 2019, San Francisco, CA: Building an African American prostate cancer biospecimen repository – case study to evaluate DNA based biomarkers of recurrence.

Presentation #5 University of Michigan, Department of Biostatistics, 2019, Ann Arbor, MI: Prostate cancer recurrence risk prediction in African Americans using tumor genomic and histopathologic image features.

i. Journal publications.

Nothing to report

ii. **other non-periodical, one-time publications.**

Nothing to report

iii. **Other publications, conference papers, and presentations.**
Nothing to report

b. **Website(s) or other Internet site(s)**

Nothing to report

c. **Technologies or techniques**

Nothing to report

d. **Inventions, patent applications, and/or licenses**

Nothing to report

e. **Other Products**

Nothing to report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

a. What individuals have worked on the project?

Name:	*Albert M. Levin, PhD*
Project Role:	*co-PI*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*2*
Contribution to Project:	*Dr. Levin is the PI for the Henry Ford site. In addition to the design of the study, he is overseeing the process of tissue acquisition, pathology review, clinical/pathological data abstraction, histological staining and sectioning of the blocks, specimen shipment, data analysis, and manuscript writing.*
Funding Support:	*DoD; The Fund for Henry Ford*

Name:	*Pamela L. Paris, PhD*
Project Role:	*co-PI*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*3*
Contribution to Project:	*Dr. Paris is the PI for the UCSF site, which is doing all of the DNA extractions and copy number array profiling. She is also working closely with Dr. Levin on oversight of pathologic review and tissue preparation, as well as development and writing of manuscripts based on the cohort.*
Funding Support:	*DoD*

Name:	*Sudha Sadasivan, PhD, MPH*
Project Role:	*Study coordinator*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*2*
Contribution to Project:	*Dr. Sadasivan is responsible for the day-to-day management of all aspects of the project.*
Funding Support:	*DoD; The Fund for Henry Ford*

Name:	*Indra Adrianto, PhD*
Project Role:	*Bioinformatician*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*1*
Contribution to Project:	*Dr. Adrianto has particular expertise in the analysis of RNA sequencing data and is responsible for the analyses linking CpG sites to the expression of genes using the prostate TCGA tumors, as well as the race-ethnicity specific methylation based gene expression prediction models.*
Funding Support:	*DoD; The Fund for Henry Ford*

Name:	*Cara Canella, MS*
Project Role:	*Biostatistician*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*4*
Contribution to Project:	*Ms. Canella is a biostatistician who carried out the analyses evaluating the use of the Illumina EPIC array data to call copy number alterations. She also carried out the preliminary association analysis between copy number alterations and biochemical recurrence. Finally, she conducted all of the analyses evaluating the associations between existing clinical pathologic features and biochemical recurrence and the differences in these association by race in the Henry Ford radical prostatectomy cohort.*
Funding Support:	*DoD; The Fund for Henry Ford*

Name:	*Ian Loveless*
Project Role:	*Biostatistician*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*3*
Contribution to Project:	*Mr. Loveless is a biostatistician who performed all of the data cleaning and quality control for the Illumina EPIC. He also performed all of the single CpG site association analyses as well as the multi-CpG biochemical recurrence predictive modelling to address Aims 1 and 2, respectively.*
Funding Support:	*DoD; The Fund for Henry Ford*

b. **Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?**

New Other Support for Albert Levin:

Title:	Leveraging electronic medical records to perform large-scale diabetes pharmacogenomics among ancestrally diverse patient populations
Effort:	1.5%
Supporting Agency:	NIH-NIDDK
Grants Officer:	Christina Coriz, Grants Management Officer, NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES
Performance Period:	04/01/17 – 03/31/22
Funding Amount:	$47,988
Project Goals:	The overarching goal of this project is to use electronic medical and pharmaceutical records from a large, diverse patient population in metropolitan Detroit to identify genes involved in the glucose-lowering effects of medications used in the treatment of diabetes. In addition to identifying these genes, this application will also determine whether the variants identified are unique or shared between African American and European American individuals.
Specific Aims:	Aim 1) To create a large cohort of individuals identified through the EMR as having diabetes and being treated with metformin. Blood and saliva samples will be collected for genomic DNA, and where needed, we will obtain HbA1c and fasting blood glucose levels to supplement missing follow-up tests. The analytic goal of this aim is to assess for differences in metformin treatment response by race-ethnicity and genetic ancestry. Aim 2) To perform genome wide genotyping on the discovery set using customized arrays with increased coverage for race-ethnic population groups. The analytic goal of this aim is to use gene-based association to identify both shared and population group specific gene x metformin interactions influencing HbA1c levels. Aim 3) To replicate Aim 2 findings in a

	separate set of individuals using customized arrays with enhanced coverage of the promoted genes. Exploratory analyses will include assessing the promoted genes for gene x metformin interactions on time-to-vascular events and the occurrence of adverse drug reactions. The primary analytic goal of this aim is to replicate associations identified in Specific Aim 2 for HbA1c.
Overlap:	None

Title:	Treatment Utilization Before Suicide
Effort:	5%
Supporting Agency:	NIMH
Grants Officer:	Rita Sisco, Grants Management Officer, NATIONAL INSTITUTE OF MENTAL HEALTH
Performance Period:	03/15/2015-02/29/2020 (NCE)
Funding Amount:	$548,847
Project Goals:	This project uses data from eight Mental Health Research Network affiliated health systems across the U.S. to investigate the association between other, non-psychiatric clinical factors, including medical diagnoses, medications, procedures, and types of visits and suicide risk
Specific Aims:	Specific Aim 1: Identify the types and timing of clinical factors prior to suicide. We will use VDW clinical codes to identify the occurrence of defined, general medical diagnoses prior to suicide. The selected factors are common, burdensome diseases often observed in practice, which have been linked with suicide in other populations. We will document the frequency, proportion, and timing of factors prior to suicide. Specific Aim 2a: Compare clinical factors before suicide to a matched sample of health care users. This hypothesis-testing matched case-control study will compare group variation in clinical factors studied in Aim 1 and preliminary studies. We will match on age, sex, site, and treatment period (year). This Aim assesses whether factors differed between suicide cases and controls using a conditional logistic regression. Specific Aim 2b: Detect associations between additional clinical factors and suicide. This hypothesis-generating matched case-control study will use environment-wide association study (EWAS) methods to detect variation in additional clinical factors (other diagnoses, medications, procedures) between suicide cases and controls. Then, latent class modeling will be used to group correlated factors into profiles or dimensions of risk, so we can investigate associations between those profiles and suicide. Specific Aim 3: Develop a prediction model of clinical factors prior to suicide. We will use significant factors identified in Aims 2a-2b to develop a prediction model of clinical factors and suicide risk. This innovative approach will study interactions between factors and for 'mental health need.' Specific Aim 4: Investigate indicators of 'hidden' mental health need in general medical chart notes prior to suicide. In this Exploratory Aim, we will abstract medical chart notes from a sample of individuals who died by suicide and made general medical visits at HFHS prior to death, but did not have a coded 'mental health need.' This Aim will 1) determine whether 'mental health needs' were present in general medical chart notes, but not documented with official codes, and if so, 2) use terms to develop a Natural Language Processing (NLP) algorithm to detect risk in clinical text notes.
Overlap:	None

Title:	Trans-America Consortium of the Health Care Systems Research Network for the Precision Medicine Initiative Cohort Program
Effort:	5%
Supporting Agency:	NIH
Grants Officer:	Irene Haas, Agreement Officer, OFFICE OF THE DIRECTOR, NATIONAL INSTITUTES OF HEALTH
Performance Period:	04/01/2018-03/31/2023
Funding Amount:	$10,639,769
Project Goals:	This project seeks to establish TACH as a Healthcare Provider Organization recruitment site seeking to enroll 10,000 participants in Year 1 for the Precision Medicine Initiative Cohort Program.
Specific Aims:	Aim 1) to recruit and retain 150,000 participant across the Health Care Systems Research Network; Aim 2) Engage participants through the use of patient liaisons to determine their views on genetic testing.
Overlap:	None

Title:	The Role of Gut Microbiota-Derived Exosomes in Neurovascular Unit After Stroke
Effort:	5%
Supporting Agency:	American Heart Association
Grants Officer:	American Heart Association
Performance Period:	07/01/2018-06/30/20
Funding Amount:	$90,909
Project Goals:	This study is exploring whether stroke induces changes in the bacteria trafficked in gut derived exosomes and whether these changes impact functional recovery.
Specific Aims:	Aim 1: To test the hypothesis that acute stroke influences the composition of gut microbiota-derived exosomes, thereby contributing to neurovascular unit damage and neurological outcome. Aim 2: To test the hypothesis that exosomes derived from microbiota dysbiosis result in neurovascular unit impairment via TLR4-mediated signaling after stroke.
Overlap:	None

Title:	Characterization of European American and African American Sarcoidosis via Immunogenetics
Effort:	3%
Supporting Agency:	NIH
Grants Officer:	Benjamin Sakovich, Grants Management Officer, NATIONAL HEART, LUNG, AND BLOOD INSTITUTE
Performance Period:	04/01/2019-03/31/2023
Funding Amount:	$20,000
Project Goals:	This study is utilizing single cell sequencing of blood based immune cells to characterize similarities and differences between European Americans and African Americans with sarcoidosis and to identify underlying immunophenotypes of the this heterogeneous disease.
Specific Aims:	The aims of this project are 1) to identify novel candidate genes and their most likely causal variants to be investigated in future mechanistic studies of sarcoidosis susceptibility and persistence via differential expression (DE) analysis of RNA sequencing data of whole blood from both EA and AA sarcoidosis patients and

	matched controls; 2) define the cell subsets in which the most likely causal variants operate and identify novel cell-specific effects by DE analysis of RNA sequencing data from blood-derived single cells in our cohort of EA and AA cases and matched controls; 3) replicate our list of candidate genes, their most likely causal variants and the specific cell type in which they influence disease using blood, tissue and single cells from a cohort of patients recruited based on the demographic and clinical criteria that are associated with our causal variants.
Overlap:	None

Title:	Microbiota and Allergic Asthma Precision Prevention (MAAP2)
Effort:	2%
Supporting Agency:	NIAID
Grants Officer:	NIAID
Performance Period:	09/05/2019-08/31/2023
Funding Amount:	$2,166,555
Project Goals:	This program project is focused on the maternal and infant's microbiotas and their effects on immune development, bolstered by parallel murine studies.
Specific Aims:	Project 1) What maternal characteristics, life-style factors, and exposures during pregnancy and which environmental exposures and feeding practices after birth influence child gut microbial development? What maternal and child gut microbiotas are associated with a very low or a very high risk of allergic asthma phenotype at 2 years of age?; Project 2) In a cohort of pregnant women with current allergic asthma, do modifiable maternal exposures and life-style factors during pregnancy influence changes in prenatal maternal microbiota and transfer of specific maternal microbial organisms to the founding population of the child's gut microbiota? Are patterns of maternal microbial communities at different time points during in utero development and infant gut microbial development in the first month of life related to IgE development and sensitization at 24 months of age?; Project 3) What microbial metabolites in an infant's gut are most strongly associated with development of multi-sensitized allergic asthma versus lack of sensitization and asthma at 10 years of age? What specific microbial strains and gene functions related to these metabolites, identified through shotgun metagenomic sequencing, are transferred from the mother to child to establish the child's founding microbial gut community, and can a preventive microbial consortia be developed that is effective in a mouse model?; Project 4) How does feeding live microbes to female mice before mating alter the immune development and function of their offspring to reduce allergen and virus-induced lung inflammation? What are the relative contributions of in utero environment and breast milk to immune development? In mouse models of allergen and respiratory syncytial virus-induced inflammation, will selected consortiums of bacteria identified as protective in the above projects also exhibit protective effects in mice? Will the joint effects of bacterial consortia produce substantially greater protective effects than single organisms?
Overlap:	None

c. **What other organizations were involved as partners?**

Nothing to report

8. Special Reporting Requirements

Nothing to report

9. Appendices

Nothing to report