AWARD NUMBER: W81XWH -15 -1-0638

TITLE: Identification of Prostate Cancer Predisposition Genes on the Y Chromosome

PRINCIPAL INVESTIGATOR: Dr. Lisa Albright

CONTRACTING ORGANIZATION: University of Utah
Salt Lake City, UT 84112

REPORT DATE: October 2017

TYPE OF REPORT: Annual

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2017	Annual	21 Sep 2016 - 20 Sep 2017

4. TITLE AND SUBTITLE

Identification of Prostate Cancer Predisposition Genes on the Y Chromosome

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-15-1-0638

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Lisa A. Cannon-Albright, Craig Teerlink, Alun Thomas

E-Mail: lisa.albright@utah.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

University of Utah
Department of Internal Medicine
391 Chipeta Way, Suite D
Salt Lake City, UT 84108-1266

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

We used various methods to perform a test for an excess of prostate cancer among males sharing the same Y chromosome. Method ology is being refined by comparing 2 competing methods and selecting the preferred model. This analysis identified a ranked list of Y chromosomes from the UPDB resource.

A set of 20 samples representing the top 10 high risk and 10 low-risk Y chromosomes for genotyping with a set of 16,000 Y SNPs, and Y chromosome full genome sequencing was submitted.

We used available Utah data for ~1,000 Y chromosome SNPs on 80 high risk Y chromosomes and 150 low risk Y chromosomes with some Y chromosome genotype data available. The set of ~1,000 SNPs was used to perform a phylogenetic analysis of the high vs low r isk Y chromosomes; some clustering of high risk Y chromosomes was noted.

Analysis of 10 high risk Y sequence data compared to 10 low risk Y sequence data, identified 3 coding and 3 non -coding candidate genes/variants seen in excess in high-risk Y chromosomes and not observed in the low risk set.

We used the ranked list of Y chromosomes from our initial analysis to select a set of 100 YIDs to submit for complete Y chromosome sequencing for analysis as cases. Complete Genomics and provided us with Y chromosome seq uence data for 1,800 control men.

15. SUBJECT TERMS

Y Chromosome, case/control association analysis, phylogenetic tree

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	13	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

1. INTRODUCTION

Published evidence suggests that genes present on the Y chromosome may be involved in increasing risk for prostate cancer. We have performed analysis of independent Y chromosomes in a unique computerized genealogical data resource with linked population-wide cancer data (the Utah Population Data Base or UPDB), showing that there exist Y chromosomes associated with elevated prostate cancer risk. Investigation of the Y chromosome is challenging due to lack of recombination, which prevents classical linkage-mapping, and high content of repetitive and ampliconic sequences, which make sequencing difficult. We hypothesize the existence of Y chromosome genes/variants related to increased risk, and propose multiple genetic analyses to efficiently identify the responsible Y chromosome genes or variants.

2. KEYWORDS

Y chromosome
Prostate cancer
High-risk pedigree
Phylogeny

3. ACCOMPLISHMENTS

The PI is reminded that the recipient organization is required to obtain prior written approval from the USAMRAA Grants Officer whenever there are significant changes in the project or its direction.

• What were the major goals and objectives of the project?

Specific Aims:

Aim 1. Genotype 1 carrier for each of 300 already sampled high-risk Y chromosomes (cases) and genotype 1 sampled carrier from each of 500 independent non-high-risk Y chromosome (controls) with over 1,400 Y chromosome SNPs and 23 MicroSatellite Repeat (MSR) markers to provide dense genotypes and haplotypes for the complete Y chromosome, allowing discrimination of independent Y chromosomes.

Aim 2. Perform association analysis for the Y chromosome SNPs and MSR haplotypes in the 300 cases and 500 controls to efficiently identify markers/regions of interest on the Y chromosome.

Aim 3. Perform whole-genome sequencing of the Y chromosome for 1 sampled Y-chromosome carrier from the 20 highest-risk YID groups; initially prioritize bioinformatics analysis in the regions of interest from Aim 2.

Aim 4.a. Replicate *association findings* in a high-risk familial prostate cancer data set of 2,000 cases and 2,000 matched controls that have Illumina human 5M + ExomeChip genotypes from an active CIDR GWAS prostate cancer grant (Cannon-Albright, PI), stratify by presence of male to male transmission in family of cases.

4.b. Replicate *sequence results* in collaboration with the International Consortium for Prostate Cancer Genetics set of 600 sequenced familial prostate cancer cases; stratify by male to male transmission in family of cases.

• What was accomplished under these goals?

We used various methods to perform a test for an excess of prostate cancer among males sharing the same Y chromosome; this is necessary to allow for the possibility of autosomal segregation that looks like Y segregation. We have continued to refine this methodology,

comparing 2 competing methods and selecting the preferred model. This analysis identified a ranked list of Y chromosomes from the UPDB resource.

Using this ranked list we submitted a set of 20 samples representing the top 10 high risk and a set of 10 low-risk Y chromosomes for genotyping with a set of 16,000 Y SNPs, and for Y chromosome full genome sequencing.

We used available Utah data for ~1,000 Y chromosome SNPs on a set of 80 high risk Y chromosomes and a set of 150 low risk Y chromosomes with some Y chromosome genotype data available. We identified increased RR for prostate cancer in the previously reported region of TSPY on Yp11.2. We also used this set of ~1,000 SNPs to perform a phylogenetic analysis of the high vs low risk Y chromosomes and noted some clustering of high risk Y chromosomes.

Analysis of the 10 high risk Y sequence data compared to the 10 low risk Y sequence data identified a set of 3 coding and 3 non-coding candidate genes/variants seen in excess in high-risk Y chromosomes and not observed in the low risk set.

We used the ranked list of Y chromosomes from our initial analysis of all Utah males to select a set of 100 YIDs to submit for complete Y chromosome sequencing for analysis as cases. Complete Genomics and provided us with Y chromosome sequence data for 1,800 control men.

Data for the 100 additional samples is due in 2 months. These data will be used to statistically test the 6 candidate variants identified in the initial set of 10 high risk cases and 10 controls. In addition, we will use the new high-risk YID cases with sequence data to identify additional candidate variants observed in excess in high-risk Y chromosomes.

• **What opportunities for training and professional development did the project provide?** No formal opportunities, but we have increased our expertise with tests of association of cancer and the Y chromosome, and with with specific Y chromosome sequence data and phylogenetic analysis.

• **How were the results disseminated to communities of interest?** Data is still being received and analyzed and there are no results to disseminate at this time. The original findings were published (Cannon-Albright et al., 2014). We will publish results on our study design and significant findings when ready.

• **What do you plan to do during the next reporting period to accomplish the goals and objectives?** Based on the sequence data we receive on the second 100 Y chromosomes we will test our first set of 6 candidate genes for association, and we will use the new high-risk case data to identify a larger set of candidates. We will refine the Y risk-haplotypes identified and refine the phylogenetic assignment of high- and low-risk Y chromosomes. Y chromosome genotype and sequence data is not widely available, but we will identify collaborators with samples that can be identified as high- or low-risk for the Y chromosome and partner to generate appropriate data and to advance functional confirmation of all candidates.

What were the major goals of the project?

Aim 1. Genotype 1 carrier for each of 300 already sampled high-riskY chromosomes (cases) and genotype 1 sampled carrier from each of 500 independent non-high-risk Y chromosome (controls) with over 1,400 Y chromosome SNPs and ~20 MicroSatellite Repeat

(MSR) markers to provide dense genotypes and haplotypes for the complete Y chromosome, allowing discrimination of independent Y chromosomes.

TASK	DESCRIPTION	MONTHS
1	get IRB approval for genetic analysis of Y chromosomes	1-4
2	identify stored DNA for 300 cases and 500 controls in lab	5-7
3	perform quality control on 800 DNA samples	8-9
4	prepare plates for genotyping at the core	10
5	Use high density SNP platform and Y-STR kit to genotype 300 informative cases and 500 controls (1,400 Y chromosome SNPs and 23 Y-STR loci)	11-14

Aim 2. Perform association analysis for the Y chromosome SNPs and MSR haplotypes in the 300 cases and 500 controls to efficiently identify markers/regions of interest on the Y chromosome.

TASK	DESCRIPTION	MONTHS
7	use Plink and Gemma software to perform case control association analysis of 300 cases and 500 controls	15
8	define regions of interest for Y chromosome sequencing based on association analysis	15

Aim 3. Perform whole-genome sequencing of the Y chromosome for 1 sampled Y-chromosome carrier from the 20 highest-risk YID groups; initially prioritize bioinformatics analysis in the regions of interest from Aim 2.

TASK	DESCRIPTION	MONTHS
9	Identify 20 highest-risk YID groups with multiple prostate cancer cases sampled.	15
10	Perform whole genome sequencing on the 20 males representing the highest risk YIDs	16-18
11	Perform bioinformatics analysis of the 20 males using TOMATO pipeline and VAAST software	18-20
	Confirm variants of interest with Sanger sequencing in at least 2 carriers of each Y chromosome of interest	21-22
12	write manuscript	23-24

Aim 4.a. Replicate *association findings* in a high-risk familial prostate cancer data set of 2,000 cases and 2,000 matched controls that have Illumina human 5M + ExomeChip genotypes from an active CIDR GWAS prostate cancer grant (Cannon-Albright, PI), stratify by presence of male to male transmission in family of cases..

TASK	DESCRIPTION	MONTHS
13	Extract Y chromosome SNPs from CIDR data prostate cancer cases	16

14	identify ICPCG/CIDR prostate cancer cases with segregation consistent with Y chromosome association and stratify into 2 sets of cases (consistent with Y segregation and not)	17
15	perform case/control association analysis with SNPs identified in Aim 2 for confirmation	18-20
16	write manuscript	21-24

4.b. Replicate *sequence results* in collaboration with the International Consortium for Prostate Cancer Genetics set of 600 sequenced familial prostate cancer cases; stratify by male to male transmission in family of cases.

TASK	DESCRIPTION	MONTHS
17	extract Y chromosome exome sequence data from ICPCG sequence data for 600 familial cases	18
18	identify ICPCG sequenced cases with segregation consistent with Y chromosome association and stratify	19-20
19	identify variants of interest in genes identified in Aim 3 in ICPCG cases and confirm association with increased risk in carriers	21
20	write manuscript	21-24

What was accomplished under these goals?

Aim 1. Genotype 1 carrier for each of 300 already sampled high-risk Y chromosomes (cases) and genotype 1 sampled carrier from each of 500 independent non-high-risk Y chromosome (controls) with over 1,400 Y chromosome SNPs and ~20 MicroSatellite Repeat (MSR) markers to provide dense genotypes and haplotypes for the complete Y chromosome, allowing discrimination of independent Y chromosomes.

TASK	DESCRIPTION	MONTHS
1	get IRB approval for genetic analysis of Y chromosomes	1-4
2	identify stored DNA for 300 cases and 500 controls in lab	5-7
3	perform quality control on 800 DNA samples	8-9
4	prepare plates for genotyping at the core	10
5	Use high density SNP platform and Y-STR kit to genotype 300 informative cases and 500 controls (1,400 Y chromosome SNPs and 23 Y-STR loci)	11-14

Pilot analysis of Utah Y chromosomes with correction for multiple testing identified high-risk Y chromosomes. Pilot analysis of available MSR and SNP data for the Y chromosome showed that most available genomewide commercial products do not provide informative Y chromosome

genotypes, so a test set of high- and low-risk Y chromosome samples were sent for QC for genotype and sequence data to determine power of available products.

Task 1; IRB approval was accomplished.
Task 2; identification of DNA for high- and low-risk Y chromosomes was completed.
Task 3; QC was performed on a set of 20 pilot samples
Task 4; plates were prepared for shipment to service facilities for genotyping and sequencing
Task 5; we received data on genotyping of 16,000 Y SNPs and Y chromosome whole genome sequencing for n=10 cases and n=10 controls.

Aim 2. Perform association analysis for the Y chromosome SNPs and MSR haplotypes in the 300 cases and 500 controls to efficiently identify markers/regions of interest on the Y chromosome.

TASK	DESCRIPTION	MONTHS
7	use Plink and Gemma software to perform case control association analysis of 300 cases and 500 controls	15
8	define regions of interest for Y chromosome sequencing based on association analysis	15

Using available genotype data for the Y chromosome for Utah subjects from other studies association analysis was performed for 83 highest-risk genotyped Y chromosomes and for 154 low-risk Y chromosomes; of a total of 1030 Y SNPs from illumina OmniExpress genotypes, 242 failed missingness and 512 were monomorphic, so 277 Y chromosome SNPs were analysed. No SNPs showed significant evidence for association for high versus low risk Y chromosomes. Two SNPs were borderline significant ($p<0.07$), one at Yq11.21 and one at Yp11.2 (previously reported as associated with increased prostate cancer risk). Using additional males with genotype data from other studies we are identifying additional cases and controls to repeat this analysis with more power for publication.

Aim 3. Perform whole-genome sequencing of the Y chromosome for 1 sampled Y-chromosome carrier from the 20 highest-risk YID groups; initially prioritize bioinformatics analysis in the regions of interest from Aim 2.

TASK	DESCRIPTION	MONTHS
9	Identify 20 highest-risk YID groups with multiple prostate cancer cases sampled.	15
10	Perform whole genome sequencing on the 20 males representing the highest risk YIDs	16-18
11	Perform bioinformatics analysis of the 20 males using TOMATO pipeline and VAAST software	18-20
	Confirm variants of interest with Sanger sequencing in at least 2 carriers of each Y chromosome of interest	21-22

12	write manuscript	23-24

10 Y chromosomes are significant for excess risk after *strict multiple testing correction*. DNAs representing the top 10 high-risk Y chromsomes and 10 low-risk Y chromosomes were submitted for whole Y chromosome sequencing as a pilot test; data was received. Analysis of case compared to control sequence data identified 3 coding and 3 non-coding variants in significant excess in cases compared to low risk controls.

Aim 4.a. Replicate *association findings* in a high-risk familial prostate cancer data set of 2,000 cases and 2,000 matched controls that have Illumina human 5M + ExomeChip genotypes from an active CIDR GWAS prostate cancer grant (Cannon-Albright, PI), stratify by presence of male to male transmission in family of cases..

TASK	DESCRIPTION	MONTHS
13	Extract Y chromosome SNPs from CIDR data prostate cancer cases	16
14	identify ICPCG/CIDR prostate cancer cases with segregation consistent with Y chromosome association and stratify into 2 sets of cases (consistent with Y segregation and not)	17
15	perform case/control association analysis with SNPs identified in Aim 2 for confirmation	18-20
16	write manuscript	21-24

Permission for access to ICPCG data was approved and data was received for 2,000 cases and 1,200 controls from CIDR. ICPCG was not able to provide data on evidence for Y chromosome segregation in the prostate cancer cases provided; ICPCG-studied families are not tracked by evidence for Y-chromosome sharing to allow discrimination of high- and low-risk Y chromosomes/pedigrees. A case/control association analysis of all Y chromosome markers for *prostate cancer status (only*) identified no significant associations, as would be expected for this phenotype.
The CIDR data set had significant missing Y genotypes (~40%) making analysis of the Utah subset of data (for which high- and low-risk Y chromosomes could be identified) not possible. The Utah genotypes will be included in the association test in Aim 2.

4.b. Replicate *sequence results* in collaboration with the International Consortium for Prostate Cancer Genetics set of 600 sequenced familial prostate cancer cases; stratify by male to male transmission in family of cases.

TASK	DESCRIPTION	MONTHS
17	extract Y chromosome exome sequence data from ICPCG sequence data for 600 familial cases	18
18	identify ICPCG sequenced cases with segregation consistent with Y chromosome association and stratify	19-20
19	identify variants of interest in genes identified in Aim 3 in ICPCG cases and confirm association with increased risk in carriers	21

20	write manuscript	21-24

The Y chromosome sequence data provided by the ICPCG sequencing project at Mayo was of incredibly low quality, with coverage less than 5X, so no bioinformatics analysis was possible. In addition, ICPCG was unable to provide data on male to male transmission in the families of the cases.

We therefore identified 100 Utah Y chromosomes with a significant excess of Y-chromosome-sharing prostate cancer cases ($p<0.01$ without multiple testing correction) and performed QC and submitted these DNA samples for whole Y genome sequencing; we are awaiting return of data.

What opportunities for training and professional development has the project provided?
No formal opportunities, but we have increased our expertise with UPDB analysis of Y chromosome inheritance, and with analysis of Y chromosome sequence data, and phylogenetic analysis.

How were the results disseminated to communities of interest?
Data is still being received and analyzed and there are no results to disseminate at this time. We will publish results on our study design and significant findings.

What do you plan to do during the next reporting period to accomplish the goals?
Based on the sequence data we receive on the second 100 Y chromosomes we will test our first set of 6 candidate genes for association, and we will use the new high-risk case data to identify a larger set of canddiates. We will refine the Y risk-haplotypes identified and refine the phylogenetic assignment of high- and low-risk Y chromosomes. Y chromosome genotype and sequence data is not widely available, but we will identify collaborators with samples that can be identified as high- or low-risk for the Y chromosome and partner to generate appropriate data and to advance functional confirmation of all candidates.

4. IMPACT

Nothing to report yet. Identification of Y chromosome predisposition genes/variants will be accomplished. Knowledge of prostate cancer etiology will be advanced; specific risk calculations for some men will be possible.

What was the impact on the development of the principal discipline(s) of the project?
We have established use of the UPDB to consider hypothese of Y chromosome inheritance and developed analysis tools.

What was the impact on other disciplines?
The methods and tools developed can be extended to other phenotypes to allow investigation of the Y chromosome (which has been implicated for many cancers).

What was the impact on technology transfer?
Nothing to report.

What was the impact on society beyond science and technology?
Nothing to report yet.

5. CHANGES/PROBLEMS

Nothing to report.

Changes in approach and reasons for change

Y chromosome genetic data has not been paid much attention and existing genetic (both genotype and sequence) data was found to be of very low quality and quantity. As we discovered the lack of Y chromosome data for validation we had to revise to extract more data from our own resources. As sequencing methods advanced it became more appropriate to go directly to sequencing samples of interest rather than genotyping.

Actual or anticipated problems or delays and actions or plans to resolve them

Y chromosome sequencing is extremely time consuming and delivery of data was much slower than anticipated.

Changes that had a significant impact on expenditures

None

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

None

6. PRODUCTS

Nothing to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	**Lisa A. Albright**
Project Role:	PI/PD
Researcher Identifier:	0000-0003-2602-3668
Nearest person month worked:	1.2
Contribution to Project:	Dr. Cannon-Albright has directed the course of the research, selection of cases, and methods of analysis, as well as supervised the day to day activities of the research team.
Funding Support:	N/A

Name:	**Craig C. Teerlink**
Project Role:	Co-Investigator
Researcher Identifier:	0000-0002-1992-2326
Nearest person month worked:	1.2

Contribution to Project: Dr. Teerlink has been responsible for oversight of data quality control and genetic analyses (including association analysis and bioinformatics analysis of sequence data) and method development/testing

Funding Support:

Name: **Alun Thomas**
Project Role: Co-Investigator
Researcher Identifier: 0000-0001-5650-7044
Nearest person month worked: .83
Contribution to Project: Dr. Thomas has provided statistical expertise for comparison of Y chromosome haplotypes and graphical modeling methods. He developed models to best identify the Y chromosome specific risk for prostate cancer.
Funding Support: N/A

Name: **Steven Backus**
Project Role: Computer Professional
Researcher Identifier: N/A
Nearest person month worked: 1.5
Contribution to Project: Mr. Backus has performed data management, data extraction, security, and data storage.
Funding Support: N/A

Name: **James Farnham**
Project Role: Applied Biostatistician
Researcher Identifier: 0000-0002-8213-949X
Nearest person month worked: 1.7
Contribution to Project: Mr. Farnham has acted as the main analyst for the project, and has been responsible for the generation and quality control of all data files for analysis and for analysis of various methods testing for risk for prostate cancer by Y chromosome.
Funding Support: N/A

Name: **Kim Nguyen**
Project Role: Laboratory Specialist
Researcher Identifier: N/A
Nearest person month worked: 2.4
Contribution to Project: Mr. Nguyen has been responsible for testing the concentration and quality of DNA, preparing stored samples, performing quality controls for all samples, and the appropriate storage and inventory and shipping of all biospecimen samples.
Funding Support: N/A

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Some grants have ended and others have been started; overall FTE on this project has not been impacted.

What other organizations were involved as partners?
Organization Name: NIH funded International Consortium for Prostate Cancer Genetics (ICPCG)

Location of Organization: The ICPCG is a working group of investigators, consisting of 16 independent groups from 21 institutions from North America, Europe and Australia. The awardee of the NIH grant is located in Rochester, MN.

Partner's contribution to the project: The ICPCG has provided data.

8. APPENDICES