Award Number: W81XWH-13-1-0336

TITLE: Whole Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition

PRINCIPAL INVESTIGATOR: Mary-Claire King, PhD

CONTRACTING ORGANIZATION: University of Washington
Seattle, WA, 98195

REPORT DATE: December 2015

TYPE OF REPORT: Final

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

DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.
Whole Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition

Mary-Claire King, PhD
Tom Walsh, PhD,

Email: mcking@uw.edu

University of Washington
Seattle, WA, 98195

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

Approved for public release; distribution unlimited

As genes for inherited disease are increasingly well characterized by next generation sequencing approaches, it is clear that some mutations may act through promoters, enhancers, and other non-coding regulatory regions. Our hypothesis for this proposal is that much of the substantial remaining familial risk of breast cancer is due to a large number of individually rare alleles of moderate-to-severe effect located in the non-coding regions of the genome. For this proposal we will evaluate 30 large, extended kindreds severely affected with breast cancer, each of whom has been comprehensively evaluated in our lab by targeted genomic sequencing for mutations of all classes in all known breast cancer genes and by whole exome sequencing for coding region mutations genome-wide. These families are a unique discovery series for identification of regulatory mutations that may reveal new mutational mechanisms for breast cancer predisposition.

Genome sequencing, mutation, breast cancer susceptibility genes
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>4</td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>4</td>
</tr>
<tr>
<td>ACCOMPLISHMENTS</td>
<td>4-6</td>
</tr>
<tr>
<td>IMPACT</td>
<td>7</td>
</tr>
<tr>
<td>CHANGES/PROBLEMS</td>
<td>8</td>
</tr>
<tr>
<td>PRODUCTS</td>
<td>8</td>
</tr>
<tr>
<td>PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS</td>
<td>8-9</td>
</tr>
<tr>
<td>SPECIAL REPORTING REQUIREMENTS</td>
<td>9</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>9-10</td>
</tr>
</tbody>
</table>
1. INTRODUCTION: Despite tremendous advances in mutation detection with gene panels and exome sequencing, most families severely affected by breast cancer do not have causative alleles identified from the protein-coding region of the genome. We hypothesize that the causal mutations in many of these families lie in uncharacterized regulatory regions of the genome. Through whole genome sequence analysis of severely affected families and functional annotation and experimental evidence, we undertook to identify new mutational mechanisms that predispose to breast cancer. Our ultimate goal is to enable information on newly identified mutations and mutational mechanisms to be useful to clinicians and to women and their families.

2. KEYWORDS: Breast cancer, BRCA1, BRCA2, whole genome sequencing, promoter, enhancer, transcription factor binding site, gene regulation, mutation.

3. ACCOMPLISHMENTS:

What were the major goals of the project?

The major goals of this project as stated in the approved SOW are listed below:

TASK 1. (Walsh) Perform whole genome sequencing of germline DNA from 100 breast cancer patients selected from 30 severely affected families.

1a. Prepare 100 standard paired end library with 300-400bp inserts (months 1-3)
1b. Prepare 100 mate-paired library with tightly defined 6kb inserts (months 1-3)
1c. Sequence the paired end and mate-paired libraries on a HiSeq2500 (months 2-9)

All components of TASK 1 have been completed.

TASK 2. (Walsh) Annotating sequencing genome variants with respect to population frequency and overlap with ENCODE regions.

2a. Align reads to the reference sequence (months 4-10)
2b. Identify SNPs, indels, CNVs and rearrangements by bioinformatics tools (months 4-10)
2c. Filter variants from Task 2b against public databases to remove common events (months 4-10)
2d. Filter rare and private variants from Task 2c within families to obtain segregating variants (months 4-10)
2e. Compare surviving events from Task 2d to ENCODE regions (months 4-10)
2f. Filter variants from Task 2e to ENCODE variants mapped only in breast tissues/lines (months 4-10)

All components of TASK 2 have been completed.

TASK 3. (King) Characterize potential regulatory variants.

3a. Generate enhancer constructs with wildtype and variant regulatory regions (months 8-16)
3b. Transfect constructs into cell lines, monitor luciferase activity (months 8-16)
3c. Measure gene expression in patients' lymphoblasts (months 8-16)

All components of TASK 3 have been completed.

TASK 4. (King) Resequence mutant regulatory regions in large series of patients to identify additional mutations.

4a. Design molecular inversion probes (MIPs) for promising regulatory regions (months 12-24)
4b. Perform MIP amplification, hybridization and sequencing (months 12-24)
4c. Annotate variants within regulatory regions with respect to frequency (months 12-24)
4d. Statistical analysis of variants (months 12-24)

All components of TASK 4 have been completed.
What was accomplished under these goals?

TASK 1: (Walsh) We prepared both library types and generated whole genome sequencing data on the 100 breast cancer patients. We obtained median sequence coverage of 34X, with 99.89% of basepairs read at least 8 times. Figure 1 below illustrates the distribution of read depths across the 100 samples and Figure 2 shows the distribution of median read depths.

Figure 1. Distribution of read depths across the 100 whole genome sequenced samples.

Figure 2. Distribution of median read depths across the 100 whole genome sequenced samples.
TASK 2: (Walsh) In Year 2 after sequencing was completed, we developed a functional annotation approach to filter variants from the whole genome sequences. Sequence reads from TASK1 were aligned to the genome with BWA\(^1\) which is presently the preferred alignment tool of the genomics community. Single nucleotide variants (SNVs) and small indels (1-50bp) we identified by GATK\(^2\), UnifiedGenotyper\(^3\), FreeBayes\(^4\) and Samtools\(^5\). Structural Variants (SVs) greater than 500bp were identified with Lumpy\(^6\), SVtyper\(^7\) and CNVnator\(^8\). We felt it was important to use all these calling tools since they have different specificities and sensitivities.

In order to remove common and likely benign SNPs, indels and SVs for follow up in TASK3 we filtered using these criteria:

1. Less than 1% frequency in the 1000 Genomes Project\(^9\) and Genomes of the Netherlands\(^10\)
2. Present in less than 4 of the breast cancer families that we sequenced in TASK1
3. Variant read proportion greater than 0.25
4. Read depth greater than 8 in at least one sample

The Table below shows variant data from Family 1041 categorized by functional effect.

**Table 1. Distribution of different types of mutations at different filtering levels in the whole genome sequencing data of two patients from a severely affected breast cancer Family 1041.**

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Shared</th>
<th>Rare</th>
<th>Excluding IBD0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>3,345,727</td>
<td>1,650,045</td>
<td>35,927</td>
<td>3,990</td>
</tr>
<tr>
<td>ncRNA</td>
<td>266,300</td>
<td>130,836</td>
<td>3,104</td>
<td>329</td>
</tr>
<tr>
<td>Up-downstream</td>
<td>72,754</td>
<td>35,055</td>
<td>865</td>
<td>98</td>
</tr>
<tr>
<td>Untranslated</td>
<td>48,884</td>
<td>23,492</td>
<td>435</td>
<td>49</td>
</tr>
<tr>
<td>Intronic</td>
<td>1,986,665</td>
<td>970,436</td>
<td>20,088</td>
<td>2,139</td>
</tr>
<tr>
<td>Missense</td>
<td>13,933</td>
<td>6,891</td>
<td>48</td>
<td>11</td>
</tr>
<tr>
<td>Silent</td>
<td>15,150</td>
<td>7,567</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>Nonsense</td>
<td>116</td>
<td>45</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Splice</td>
<td>198</td>
<td>120</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Stoploss</td>
<td>15</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>In-frame</td>
<td>385</td>
<td>162</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Frame-shift</td>
<td>199</td>
<td>92</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5,750,326</td>
<td>2,824,749</td>
<td>60,503</td>
<td>6,622</td>
</tr>
</tbody>
</table>

In order to prioritize SNPs, indels and SVs for follow up in TASK3 we filtered remaining variants by segregation in each family. We excluded any variant that was not present in at least three women with breast cancer in a family.

The remaining variants were annotated for regulatory potential as follows:

1. ENCODE regions\(^11\): We used chromatin predictions from Human Mammary Epithelial cell lines (DNasel and H3K27Ac signals). In addition, transcription factor CHiP-seq and DNasel clusters from 7 breast cancer cell lines and finally predicted Transcription Factor binding motifs.

2. Potential motif disruption: For each potential ENCODE region and overlapping Transcription Factor binding motifs, we calculated motif score changes using position weight matrices.

3. Non coding prediction: We used scores generated by DANN\(^12\), FATHMM\(^13\) and FunSeq2\(^14\) that combine population frequencies and nucleotide conservation for non-coding regions of the genome.

At the end of TASK 2 we selected 112 non-coding variants from the 33 breast cancer families that shared these
features:

(i) segregated with breast cancer in a family
(ii) private or ultra-rare in population databases
(iii) predicted to disrupt a potential regulatory element
(iv) within 2MB of a known breast cancer predisposition gene

TASKS 3 and 4 were performed in Year 2.

TASK 3. (King) We assessed gene expression levels by RT-PCR and targeted RNAseq using RNA from freshly acquired blood samples of the 33 patients harboring the 112 variants described in TASK 2. Of the 112 variants evaluated 11 showed expression profiles suggesting dysregulation of gene transcription.

TASK 4. (King) We genotyped the 11 variants from TASK3 in 960 female cancer free controls and 960 unrelated breast cancer patients different from those used in TASK1. We did not observe these variants in any of the 960 controls or in any other familial breast cancer cases confirming they are rare or potentially private to their host families.

We also sequenced approximately 300bp around each of the 11 variants of interest in 960 female cancer-free controls and 960 unrelated breast cancer patients. There were no additional rare variants that were predicted to disrupt a regulatory element.

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

Nothing to report

**How were the results disseminated to communities of interest?**

We are in the process of writing two manuscripts describing:

(i) Our approach for targeted RNA sequencing to evaluate non-coding variants
(ii) Our bioinformatics analysis of family based whole genome sequencing data

Our main findings describing the 11 interesting non-coding variants will require additional confirmation before publishing and presentation at scientific meetings.

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

Nothing to report (end of this project)

**4. IMPACT**

**What was the impact on the development of the principal discipline(s) of the project?**

Our approach integrated whole genome sequencing with experimental biology and with application and development of bioinformatics tools to discover regulatory variants that may predispose women to develop breast cancer. Through our publications and data we hope that this approach will encourage other researchers to examine non-coding portions of the genome for additional cancer predisposing mutations.

**What was the impact in other disciplines?** Nothing to report

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

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

Through our publications and data we hope that non-coding variants that definitely impact breast cancer gene expression will be incorporated into clinical care.
5. CHANGES/PROBLEMS

Changes in approach and reasons for change Nothing to report

Actual or anticipated problems or delays and actions or plans to resolve them Nothing to report

Changes that had a significant impact on expenditures Nothing to report

Significant changes in the use or care of human subjects, vertebrate animals, biohazards, and/or select agents Nothing to report

6. PRODUCTS

Publications, conference papers, and presentations

Nothing to report at present, although we have two manuscripts in preparation

Websites or other Internet sites Nothing to report

Technologies or techniques Nothing to report

Inventions, patent applications, and/or licenses Nothing to report

Other products Nothing to report

7. PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

<table>
<thead>
<tr>
<th>Name:</th>
<th>Tom Walsh, PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Role</td>
<td>PI</td>
</tr>
<tr>
<td>Research Identifier</td>
<td>24</td>
</tr>
<tr>
<td>Nearest person month worked</td>
<td>24</td>
</tr>
<tr>
<td>Contribution to Project</td>
<td>Tom Walsh performed the whole genome sequencing and participated in variant filtering and selection</td>
</tr>
<tr>
<td>Funding Support</td>
<td>NIH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name:</th>
<th>Mary-Claire King, PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Role</td>
<td>PI</td>
</tr>
<tr>
<td>Research Identifier</td>
<td>24</td>
</tr>
<tr>
<td>Nearest person month worked</td>
<td>24</td>
</tr>
<tr>
<td>Contribution to Project</td>
<td>Mary-Claire King performed candidate variant selection and supervised the expression analysis experiments</td>
</tr>
<tr>
<td>Funding Support</td>
<td>ACS, NIH, Breast Cancer Research Foundation, Komen for the Cure</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name:</th>
<th>Ming Lee, PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Role</td>
<td>Bioinformaticist</td>
</tr>
<tr>
<td>Research Identifier</td>
<td>24</td>
</tr>
<tr>
<td>Nearest person month worked</td>
<td>24</td>
</tr>
<tr>
<td>Contribution to Project</td>
<td>Ming Lee performed the whole genome sequencing data analysis and downstream bioinformatics and variant analysis</td>
</tr>
<tr>
<td>Funding Support</td>
<td>NIH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name:</th>
<th>Silvia Casadei, PhD</th>
</tr>
</thead>
</table>
Has there been a change in the active other support of the PD/PIs or senior/key personnel since the last reporting period?

No change

What other organizations were involved as partners?

None

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS

A duplicate report with the Tasks clearly marked with the responsible PI was submitted for this award.

APPENDICES

References cited:


7 Hall I. Bayesian genotyper for structural variants. https://github.com/hall-lab/svtyper


