NORTHWESTERN UNIVERSITY

Mutational Processes Modeling and Early Cancer Diagnosis

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Health and Biomedical Informatics

By

Zexian Zeng

EVANSTON, ILLINOIS

March 2019
ABSTRACT

Annual age-adjusted breast cancer incidence rates in the United States have been static for decades. More recently, the development of massively parallel, high throughput DNA sequencing has enabled the cataloging of somatic mutations in cancer. Mutations are non-random and occur within sequence motifs. These motifs provide us with evidence to infer the processes that created the mutations. Learning the patterns of mutations in these motifs, our understanding of the biological processes that generate somatic mutations in breast cancer has increased markedly. While these provide important insights into the processes responsible for somatic mutations, gaps remain, and the etiology of several signatures remains unknown. One possible explanation for the unknown etiology is due to the robust assumptions made in the signature study. Therefore, reducing the number of assumptions has a potential to decipher the unknown signature, helping us understand the signatures’ etiology.

To date, most of the motif changes were studied without insertions and deletions (INDELs) information. However, it is well known that INDELs have a large effect on the genotypes. Therefore, using whole-exome sequencing data, germline and somatic mutations are integrated and all single nucleotide variants, insertions, and deletions are interactively amalgamated as features in a deep learning model.

While great strides have been made in the treatment of breast cancer, successful prevention remains elusive. Our understanding of the mutational processes in breast cancer would ultimately improve prevention strategies. Current breast cancer prevention strategies fall into one of three categories: lifestyle modification, surgical intervention, and chemoprevention. These strategies have had, at best, limited success. In our study, we seek for biomarkers of significantly elevated breast cancer risk that can be detected in early cancer period.
Nationwide adoption of Electronic Health Records (EHRs) has given rise to a large amount of digital health data, which can be used for secondary analysis. Typical EHRs include structured data such as diagnosis codes, vitals and physiologic measurements, as well as unstructured clinical narratives such as progress notes and discharge summaries. We developed computational phenotyping to automatically mine and predict clinically significant, or scientifically meaningful phenotypes from structured EHR data, unstructured clinical narratives, or their combination.

**Dissertation Committee**

Committee chair & advisor: Yuan Luo, Ph.D.
Committee member & advisor: Seema A Khan, M.D.
Committee member & advisor: Susan E Clare, M.D., Ph.D.
Committee member: Justin Starren, M.D., Ph.D.
ACKNOWLEDGEMENTS

The pursuit of my Ph.D. study is fruitless without the involvement of many people who provide mentorship and support on multiple aspects.

This work could not have been completed without the extraordinary dedication of time, energy, coaching from my mentors Dr. Susan Clare, Dr. Seema Khan, and Dr. Yuan Luo (ordered by last name). I wanted to express my sincere gratitude for their scientific support, patience, immense knowledge, and inspiring ideas. Their guidance helped me in all the time of research and writing of my dissertation. I thank my academic advisor and committee member, Dr. Justin Starren, for his support of my training and valuable feedbacks. I am grateful to Dr. Neil Jordan, Dr. Lucy Bilaver, and Dr. Suzanne Cox, for their tremendous efforts in creating a diverse and collaborative learning environment for biomedical informatics students at the Health Sciences Integrated Program.

I would also like to acknowledge my colleagues, Andy Vo, Chengshen Mao, and Miran Choi for the time working together and for the inputs they have devoted to this dissertation. This particular doctoral candidate would never have been conceived without the encouragement, support, and help from my wife Xiaoyu Li and my parents. I am grateful for their kindness, constant support, and unconditional love.
LIST OF ABBREVIATIONS

ADE  Adverse drug event
AUC  Area under curve
BER  Base excision repair
BLCA Bladder Urothelial Carcinoma
CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma
CNNs Convolutional neural networks
COAD Colon adenocarcinoma
COGS Oncological Gene-environment Study
COSMIC Catalogue of Somatic Mutations in Cancer
CPM Counts per million
DDI Drug-drug interaction
DeepCues Deep learning for disease classification using exome sequencings
DSB Double strand breaks
DUB Deubiquitinase
EDW Enterprise Data Warehouse
EHRs Electronic Health Records
ER Estrogen receptor
ESP Exome Sequencing Project
ExAC Exome Aggregation Consortium
FDR False Discover Rate
FFPE Formalin-fixed, paraffin-embedded
GATK Genome Analysis Toolkit
GBM Glioblastoma multiforme
GBT Gradient boosted tree
GO Gene ontology
H2O2 Hydrogen peroxide
HAT Histone acetyltransferase
HM450 HumanMethylation450
HNSC Head and Neck squamous cell carcinoma
HPID Human Protein Interaction Database
ICD-9 International Classification of Disease, 9th Revision
INDELs Insertions and deletions
IRb Ionizing Radiation b
K-NN K-nearest neighbor algorithm
KIRC Kidney renal clear cell carcinoma
KIRP Kidney renal papillary cell carcinoma
KMCI  Knowledge Map Concept Identifier
LCM  Laser capture microdissection
LeakyReLU  Leaky Rectified Linear Unit
LGG  Brain Lower Grade Glioma
LIHC  Liver hepatocellular carcinoma
LOINC  Logical Observation Identifiers Names and Codes
LR  Logistic regression
LUAD  Lung adenocarcinoma
LUSC  Lung squamous cell carcinoma
MAF  Minor allele frequency
MICE  Multivariate Imputation by Chained Equations
MLP  Multiple layer perceptron
NER  Nucleotide excision repair
NLP  Natural language processing
NMEDW  Northwestern Medicine Enterprise Data Warehouse
NMF  Non-negative matrix factorization
NPCR  National Program of Cancer Registries
Nrf2  Nuclear-factor-E2-related factor 2
nsNMF  Nonsmooth nonnegative matrix factorization
OV  Ovarian serous cystadenocarcinoma
PCGP  Pediatric Cancer Genome Project
PON  Panel of Normals
POx1  Process Signature Oxidative 1
POx2  Process Signature Oxidative 2
PR  Progesterone receptor
PRAD  Prostate adenocarcinoma
PSA  Process Signature A
PSH  Process Signature H
RFC  Random forest classifier
ROC  Receiver operating characteristic
RSS  Residual Sum of Squares
SCARNAs  Cajal body-specific RNAs
SE  Standard Error
SKCM  Skin Cutaneous Melanoma
SNOMED CT  Systematized Nomenclature of Medicine-Clinical Terms
SNP  Single nucleotide polymorphism
STAD  Stomach adenocarcinoma
SVM  Support vector machine
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TCR</td>
<td>Transcription-coupled repair</td>
</tr>
<tr>
<td>THCA</td>
<td>Thyroid carcinoma</td>
</tr>
<tr>
<td>TLS</td>
<td>Translesion synthesis</td>
</tr>
<tr>
<td>UCEC</td>
<td>Uterine Corpus Endometrial Carcinoma</td>
</tr>
<tr>
<td>WHI</td>
<td>Women's Health Initiative</td>
</tr>
<tr>
<td>WTSI</td>
<td>Welcome Trust Sanger Institute</td>
</tr>
<tr>
<td>2OHE2</td>
<td>Catechol estrogens 2-hydroxyestradiol</td>
</tr>
<tr>
<td>4OHE2</td>
<td>4-hydroxyestradiol</td>
</tr>
</tbody>
</table>
DEDICATION

I dedicate this thesis to my wife and parents.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>3</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>5</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>6</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>9</td>
</tr>
<tr>
<td>DISSERTATION COMMITTEE</td>
<td>4</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>10</td>
</tr>
<tr>
<td>LIST OF FIGURES AND TABLES</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>15</td>
</tr>
<tr>
<td>I. STUDY SUMMARY AND AIMS</td>
<td>15</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
<td>19</td>
</tr>
<tr>
<td>III. INNOVATION</td>
<td>23</td>
</tr>
<tr>
<td>CHAPTER 2 COMPUTATIONAL PHENOTYPING</td>
<td>24</td>
</tr>
<tr>
<td>I. COMPUTATIONAL PHENOTYPING</td>
<td>24</td>
</tr>
<tr>
<td>II. DISTANT RECURRENCE AND NATURAL LANGUAGE PROCESSING</td>
<td>25</td>
</tr>
<tr>
<td>III. RELATED WORK</td>
<td>27</td>
</tr>
<tr>
<td>IV. METHODOLOGY</td>
<td>29</td>
</tr>
<tr>
<td>A. Cohort Description</td>
<td>29</td>
</tr>
<tr>
<td>B. Structured Clinical Data</td>
<td>30</td>
</tr>
<tr>
<td>C. Clinical Narratives</td>
<td>32</td>
</tr>
<tr>
<td>D. Feature Generation</td>
<td>33</td>
</tr>
<tr>
<td>E. Prediction Model and Evaluation</td>
<td>33</td>
</tr>
<tr>
<td>V. EXPERIMENT RESULTS</td>
<td>34</td>
</tr>
<tr>
<td>VI. DISCUSSION</td>
<td>38</td>
</tr>
<tr>
<td>VII. CONCLUSIONS</td>
<td>41</td>
</tr>
<tr>
<td>CHAPTER 3 MUTATIONAL PROCESSES IN BREAST CANCER</td>
<td>43</td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>43</td>
</tr>
<tr>
<td>II. METHODS</td>
<td>43</td>
</tr>
<tr>
<td>A. Somatic mutations</td>
<td>44</td>
</tr>
<tr>
<td>B. Signature modeling</td>
<td>44</td>
</tr>
<tr>
<td>C. Gene expression level change</td>
<td>44</td>
</tr>
<tr>
<td>D. Germline variants and somatic mutations discovery</td>
<td>45</td>
</tr>
<tr>
<td>E. Methylation discovery</td>
<td>46</td>
</tr>
<tr>
<td>F. Interaction between gene expression level and signature exposure level</td>
<td>47</td>
</tr>
<tr>
<td>III. RESULTS</td>
<td>48</td>
</tr>
<tr>
<td>A. Mutational patterns</td>
<td>48</td>
</tr>
<tr>
<td>B. Signatures A and H</td>
<td>48</td>
</tr>
<tr>
<td>C. Process Signature Ox1</td>
<td>49</td>
</tr>
<tr>
<td>D. Correlation of Process Signature Ox1 with RNA expression</td>
<td>51</td>
</tr>
<tr>
<td>E. Correlation with germline variants and methylation</td>
<td>53</td>
</tr>
<tr>
<td>F. Process Signature Ox2</td>
<td>53</td>
</tr>
<tr>
<td>G. Correlation of Process Signature Ox2 with RNA expression</td>
<td>55</td>
</tr>
<tr>
<td>H. Correlation of PSOx2 with somatic mutations</td>
<td>56</td>
</tr>
<tr>
<td>I. Interaction between genes and PSOx2 and the effect on survival</td>
<td>57</td>
</tr>
</tbody>
</table>
LIST OF FIGURES AND TABLES

Figure 1: Workflow to identify the cohort. .................................................................................................................. 29
Figure 2: Receiver operating characteristic (ROC) curves using different methods. Each experiment is replicated 20 times
using different shuffled stratified splits and all derived ROC curves are plotted. .............................................................. 37
Figure 3: Mutational patterns of Process Signature Ox1 and Process Signature Ox2. The specific nucleotide change is
outside the circles. The 5’ neighbors of the mutated nucleotide are at the inner circles. The 3’ neighbors of the
mutated nucleotide are at the far end of bars. The green circles and pink circles are drawn at 30% and 60% of the
y axis........................................................................................................................................................................ 49
Figure 4: Mutational patterns of Process Signature Ox1 and Process Signature Ox2. The specific nucleotide change is
outside the circles. The 5’ neighbors of the mutated nucleotides are at the inner circles. The 3’ neighbors of the
mutated nucleotide are at the far end of bars. The green circles and pink circles are drawn at 30% and 60% of the
y axis........................................................................................................................................................................ 51
Figure 5: The correlation between process signature exposure and RNA expression level. A. Process Signature Ox1 B.
Process Signature Ox2.................................................................................................................................................. 52
Figure 6: Data obtained from riboflavin photooxidation (ionization potential) and codon................................................................................. 56
Figure 7: Process Signature Ox2 and PAM50 subtype and triple negative status. A. PAM subtypes. B. ER, PR, and HER2
status. The numbers in y-axis are the numbers that individuals exposed to the process signature Ox2. .............................. 58
Figure 8: Kaplan-Meier survival curves for high Process Signature Ox2 versus low Process Signature Ox2. P-values
were calculated using log-rank tests. A. Tumors with high SLC25A29 expression level. B. Tumors with low
SLC25A29 expression level........................................................................................................................................ 59
Figure 9: Feature generation for proposed models (A) The transcript sequences were retrieved from RefSeq and were
formed as a consensus matrix (B) One’s germline variants were embedded in the consensus matrix, forming a
germline matrix for each sample. The brown dots are the germline variants that one has, including polymorphisms, deletions, and insertions. As an illustration, single nucleotide polymorphisms are identified and
embedded in transcript A, E, and H, in-frame shift deletion is embedded in transcript B. In-frame shift insertion is
embedded in transcript C. Frame shift deletion is embedded in D, and a frameshift insertion is embedded in
transcript E. For this sample, the transcript F and G remained the same. (C) One’s somatic mutations were
embedded in the germline matrix (from B), forming a germline and somatic matrix. The green dots are the somatic
mutations that one has, including SNVs, insertions, and deletions. As an illustration, the tissue gained somatic
mutations in transcript A, E; gained a stop loss in transcript F; and gained a deletion that shifted the frame in
transcript G................................................................................................................................................................. 66
Figure 10: The architecture of the convolutional neural network. Component (A) is the input layer, which is a one hot
encoding with the column number equals 64 (number of codons) and the row number equals the number of
codons in the transcript. Component (B) is the encoder component contains a sequence of layers, each is
consisting of a convolutional layer, followed by a Leaky Rectified Linear Unit and average pooling layer. The
number of convolution layers is determined by the gene length. Component (C) is a fully connected layer that
combined all the outputs from the component (B) and has k outputs for k diseases...................................................... 68
Figure 11: Comparing prediction accuracy between DeepCues and baseline models, including penalized logistic
regression (LR) and support vector machine (SVM) with linear kernel........................................................................... 70
Figure 12. A. Design of the BBCAR study. B. Study design illustrating the number of sequenced samples. ................. 77
Figure 13: Allele frequencies of the somatic mutations that were called in both array and whole exome sequencing ... 82
Figure 14: Comparison of the called mutations between genotype array and WES. Somatic mutations were called using
Mutect2, VarScan2, and VarDict. With different allele frequency difference cutoff, we obtained different
accuracy. The accuracies were plotted. ............................................................................................................................ 83
Figure 15: Workflow to train and test a predictive model for somatic mutation identification .................................. 84
Figure 16: Performance of different machine learning models in the holdout test. Penalized logistic regression (LR);
linear SVM; random forest classifier (RFC); gradient boosted tree (GBT); k-nearest neighbor algorithm (K-NN);
SVM with RBF kernel; multiple layer perceptron (MLP) ............................................................................................. 87
Figure 17: Comparison of the called mutations between genotype array and WES in the cancer samples in the three
validation samples. Somatic mutations were called using Mutect2 and our developed somatic mutation prediction
MODEL. WITH DIFFERENT ALLELE FREQUENCY DIFFERENCE CUTOFF, WE OBTAINED THE ACCURACY PLOT .......................................................... 88

FIGURE 18: COHORT AND CATALOG OF SOMATIC MUTATIONS IN 204 BENIGN BREAST BIOPSIES. A. MUTATION CLASSES IN THE CASE GROUP AND CONTROL GROUP B. THE TOP 10 MUTATED GENES IN THE CASE GROUP AND CONTROL GROUP. C. CATALOG OF BASE SUBSTITUTIONS, INSERTIONS/DELETIONS IN THE 204 BENIGN BREAST BIOPSIES .................................................................................................. 89


FIGURE 20: POSITION OF MUTATIONAL ALTERATIONS IN THE PROTEIN STRUCTURE OF FLG, GNAS, AND CTNNB1 ......................................... 91

FIGURE 21: SCHEMA FOR IDENTIFYING MUTATIONS WITHIN THE CASE GROUP .......................................................................................... 92

FIGURE 22: MUTATIONAL PROCESSES IDENTIFIED IN THE CASE GROUP AND THE CONTROL GROUP. BOTH GROUPS HAVE THE AGING SIGNATURE AND MISMATCH REPAIR SIGNATURE. WE ALSO IDENTIFIED A NEW SIGNATURE IN EACH GROUP .................................................. 94

FIGURE 23: THE NOVEL SIGNATURE WE HAVE IDENTIFIED, IS SIGNIFICANTLY ASSOCIATED WITH TRIPLE NEGATIVE BREAST CANCER ....... 95

FIGURE 24: RECURRENT SOMATIC COPY NUMBER VARIATION IN THE CASE GROUP .................................................................................... 96

TABLE 1: COHORT DISTRIBUTION IN THE TRAINING AND GENERALIZATION SET .............................................................................................. 30

TABLE 2: THE NAME AND CORRESPONDING CATEGORIES (VALUES) OF THE 18 RETRIEVED STRUCTURED CLINICAL VARIABLES. IDC IS INVASIVE DUCTAL CARCINOMA, DCIS IS DUCTAL CARCINOMA IN SITU, ILC IS INVASIVE LOBULAR CARCINOMA, NETWORK CATEGORY IS THE NETWORK OF PATIENT’S INSURANCE PLAN ........................................................................ 31

TABLE 3: DESCRIPTIVE SUMMARIES OF 1,995 SUBJECTS’ CLINICAL DATA. THE SIGNIFICANCE TEST IS PERFORMED BETWEEN THE RECURRENCE GROUP AND THE NON-RECURRENCE GROUP. ONLY DATA WITH P-VALUES LESS THAN 0.05 ARE PRESENTED. DR STANDS FOR DISTANT RECURRENCE. THE MEAN AND STANDARD DEVIATION ARE CALCULATED FOR CONTINUOUS VARIABLES. NUMBERS AND PERCENTAGES ARE PRESENTED FOR CATEGORICAL VARIABLES. P-VALUES ARE OBTAINED USING STUDENT’S T-TEST FOR CONTINUOUS VARIABLES AND CHI-SQUARED TEST FOR CATEGORICAL VARIABLES .................................................................................. 35

TABLE 4 THE NUMBER OF FEATURES AND THE AUC SCORES OBTAINED IN THE CROSS-VALIDATION USING 70% OF THE GROUND TRUTH DATA .............................................................................................................................................. 36

TABLE 5: THE NUMBER OF FEATURES AND THE AUC SCORES OBTAINED IN THE EXTERNAL TEST USING THE TEST SET (599 SAMPLES) ......... 36

TABLE 6: TOP 15 VARIABLES WITH THEIR CORRESPONDING COEFFICIENTS ............................................................................................... 38

TABLE 8: THE 28 GENES WITH A GENE BURDEN MOST HIGHLY ASSOCIATED WITH PSOG2. THE HIGHER THE COEFFICIENT THE GREATER THE CORRELATION. ................................................................................................................................... 57

TABLE 8: THE NUMBER OF SAMPLES OF EACH CANCER AND THE CORRESPONDING NUMBER OF GERMLINE VARIANTS AND SOMATIC MUTATIONS. VARIANTS ANNOTATED WITH MODERATE EFFECTS ARE DEFINED AS MISSENSE MUTATIONS AND INFRAME SHIFTS AND VARIANTS ANNOTATED AS HIGH EFFECTS ARE DEFINED AS NONSENSE MUTATIONS. ........................................................................................................... 64

TABLE 9: PRECISION AND RECALL FOR OUR PROPOSED MODEL. THE EXPERIMENT IS REPLICATED FOR 10 TIMES AND THE NUMBER IN PARENTHESES IS STANDARD ERROR. THE BOLDED NUMBER ARE THOSE THAT SIGNIFICANTLY IMPROVED AFTER ADDING SOMATIC MUTATION INFORMATION. .......................................................................................... 71

TABLE 10: PRECISION AND RECALL FOR OUR PROPOSED MODEL. THE EXPERIMENT IS REPLICATED FOR 10 TIMES AND THE NUMBER IN PARENTHESES STANDARD ERROR. THE BOLDED NUMBER ARE THOSE THAT SIGNIFICANTLY IMPROVED AFTER ADDING SOMATIC MUTATION INFORMATION. ........................................................................................................... 72


TABLE 12: DISTRIBUTIONS OF DEMOGRAPHIC DATA AND TUMOR CHARACTERISTICS BETWEEN THE CASE GROUP AND THE CONTROL GROUP. STUDENT’S T-TESTS WERE PERFORMED FOR CONTINUOUS VARIABLES AND PEARSON’S CHI-SQUARED TESTS WERE PERFORMED FOR CATEGORICAL VARIABLES. .......................................................................................... 79

TABLE 13: LIST OF FEATURES USED IN THE PREDICTIVE MODEL FOR SOMATIC MUTATION IDENTIFICATION. **ALLELE FREQUENCY** IS THE MUTATION ALLELE FREQUENCY, **REF DEPTH** IS THE READ DEPTH OF THE REFERENCE ALLELE, **FRE COHORT** IS THE NUMBER OF APPEARANCES OF THIS MUTATION IN THE COHORT. SNP COMMON IS A BINARY VARIABLE INDICATING THAT THE MUTATION APPEARS
IN THE DATABASE OF dbSNP AFTER REMOVING THOSE FLAGGED SNPs (SNPs < 1% minor allele frequency (MAF) (or unknown), mapping only once to reference assembly, flagged in dbSNP as "clinically associated"). ‘COSMIC’ is a binary variable indicating whether the mutation appears in the COSMIC version 80. The other features were derived from functional annotations by ANNOVAR. ................................................................. 85

TABLE 14: THE NUMBER OF MUTATIONS IN THE TRAIN SET AND TEST SET. THE GERMLINE VARIANTS WERE TREATED AS PREDICTIVE NEGATIVE AND SOMATIC MUTATIONS WERE TREATED AS PREDICTIVE POSITIVE. USING TUNED MULTIPLE LAYER PERCEPTRON MODEL TO PREDICT MUTATIONS IN THE HOLDOUT TEST SET, THE AUC SCORE, PRECISION, RECALL, AND F-MEASURE ARE REPORTED........ 87

TABLE 15: RECURRENT MUTATIONS ENRICHED IN THE CASE CONTROL ........................................................................................................... 93
CHAPTER 1 INTRODUCTION

I. Study Summary and Aims

Breast cancer is the highest incident of cancer in women in the United States; it is expected to account for 29% of new cancer diagnoses in women in the coming year.¹ Annual age-adjusted breast cancer incidence rates in the United States have been static for decades except for a modest decrease following the publication of the 2002 report of the Women's Health Initiative (WHI) trial of estrogen plus progestin, which led to a substantial decrease in the use of menopausal hormone therapy in the United States.² More recently, the development of massively parallel, high throughput DNA sequencing has enabled the cataloging of somatic mutations in cancer. Mutations are non-random and occur within sequence motifs. These motifs provide evidence from which we can infer the process that created the mutations. Learning the changes in these motifs, our understanding of the biological processes that generate somatic mutations in breast cancer has increased markedly over the past five years. About 30 “mutational signatures” have been produced.³⁻⁷ While these provide important insights into the processes responsible for somatic mutations, gaps remain, and the etiology of several signatures remains unknown. One possible explanation for the unknown etiology is due to the robust assumptions made in the signature study. Therefore, reducing the number of assumptions has a potential to decipher the unknown signature, helping us understand the signatures’ etiology. In addition, most of the motif changes were studied without insertions and deletions (INDELs) information. It is well known that INDELs have a large effect on the genotypes. Therefore, when building a model, INDEL and SNP information that falls in the same sequence need to be considered.

Current breast cancer prevention strategies fall into one of the three categories: lifestyle
modification, surgical intervention, and chemoprevention. These strategies have had, at best, limited success. Lifestyle modifications will require education, public health initiatives, and institution of public policies such as was done for tobacco eradication efforts. Surgical interventions are limited to those individuals with germline mutations in breast cancer susceptibility genes, which are non-modifiable risk factors and a fixed percentage of the population. There has been no enthusiasm for chemoprevention and the reasons for this are multifactorial including a sense of wellness or lack of symptoms being perceived as low risk, concerns about side-effects, lack of information about the causes of breast cancer and difficulties in understanding risk estimates. In order to develop prevention strategies to reduce breast cancer incidence, the biomarkers of significantly elevated breast cancer risk must be identified and validated. In addition, biomarkers identified in the cancer tissues need to be significantly different from those in the controls, that these mutations will have transcriptional consequences that directly affect processes involved with oncogenesis and that the mutations individually or in combination will have phenotypic consequences.

A phenotype is an expression of the characteristics that result from genotype variations and an organism’s interactions with its environment. Studies have analyzed a wide range of phenotypes affected by a specific genetic variant, which puts demand for automated phenotype generation. Nationwide adoption of Electronic Health Records (EHRs) has given rise to a large amount of digital health data, which can be used for secondary analysis. Typical EHRs include structured data such as diagnosis codes, vitals and physiologic measurements, as well as unstructured clinical narratives such as progress notes and discharge summaries. Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful phenotypes from structured EHR data, unstructured clinical narratives, or their combination. Developing an
automated pipeline to retrieve cancer phenotype data from EHRs and clinical narratives is necessary. We will test our questions and hypotheses experimentally with the following specific aims.

Aim 1: Identify breast cancer phenotypes using natural language processing

The emerging cancer prognosis research has directed efforts towards identifying cancer phenotypes accurately and efficiently. Manual chart review is one of the traditional methods used to identify breast cancer phenotypes. Unfortunately, chart review is a time-consuming and costly process. It limits the number of samples available for research and is not feasible for large cohort studies. Furthermore, it is subject to human error in data analysis. Computational phenotyping aims to automatically mine or predicts clinically significant, or scientifically meaningful phenotypes from structured EHR data, unstructured clinical narratives, or combination of the two.

1. Develop a system that processes and combines structured EHR data and unstructured clinical narratives as features
2. Develop models to identify breast cancer phenotypes within a cohort of breast cancer patients
3. Deliver a gold-standard data set with rich, validated information for further breast cancer research

Aim 2: Decipher breast cancer signatures

Somatic mutations are non-random and occur within sequence motifs in patterns. The motifs provide evidence from which to infer the process that created the mutation. Using this information, signatures have been introduced to describe the mutational process for breast cancer development. The processes include, for example, aging, the activity of the AID/APOBEC family of cytidine deaminases, pyrimidine dimer formation by ultraviolet light, exposure to tobacco mutagens and aristolochic acid. However, etiology of some breast cancer signatures remains unknown. General
assumptions have been made during the signature development, resulting in a high chance of mixing multiple signals into one signature. Therefore, making it difficult to explain the etiology upon one combined signature. Thus, we aimed to

1. Develop a new method in which the specific nucleotide change (e.g., C>T), the nucleotides immediately 5’ and 3’ of the mutation (e.g., 5’: C; 3’: G), and gene coding direction are all considered. The summary of these mutation characteristics forms a mutational profile for each tissue sample. Putting multiple samples’ profiles together forms a sparse matrix with the number of samples as rows and the mutation characteristics as columns. Nonsmooth nonnegative matrix factorization can then be applied to enable the discovery of intrinsic patterns in this sparse matrix.

2. Acquire, process, and analyze the samples’ matched clinical phenotypes, RNA expression levels, methylation levels (450K), and germline variants, to discover the underlying biological mechanisms for the signatures.

Aim 3: Develop a deep learning model to study mutations within sequence, interactively amalgamated somatic mutations and germline variants as features

Deep learning has recently emerged based on big data, the power of parallel computing and sophisticated algorithms are exponentially more efficient than the conventional models in learning intricate patterns from high-dimensional raw data with little guidance. The development of deep learning makes it possible to study the gene reads as a sequence instead of a single chromosomal position, in such sequences, germline variants, somatic mutations can be constructed in the sequences. When studying the sequences, SNP, INDEL, and motif information are all included in the model. Inspired by the advances of development of deep learning models and available abundant genome sequencing data, we aimed to:
1. Acquire germline variants, somatic mutations for several cancers, and construct the information into gene sequence reads.

2. Apply convolutional neural network to study the constructed sequences to learn a cancer classifier.

3. Study sequence reads (genes) in the classifier to discover potential novel cancer genes.

Aim 4: Identify biomarkers that presage the development of breast cancer

It has been suggested that the breast tissue of women destined to develop breast cancer harbor mutations within their breast tissue years. Therefore, we hypothesized that the mutations present in the benign breast biopsies of women, who went on to develop breast cancer years later, are different from those who did not develop this disease and these mutations could be used as markers of the risk of development of breast cancer. We designed a case-control study of benign breast biopsy tissues and we sequenced the DNA of retrieved tissue blocks from the benign breast biopsies of 135 patients who subsequently developed breast cancer (cases) and from 69 matched controls, who have not developed breast cancer. With the samples sequenced, we aimed to

1. Validate the quality of sequence data using SNP array.

2. Differentiate the somatic mutations and germline variants

3. Identify biomarkers that presage breast cancers using the validated sequencing data. Include repetitive somatic mutation study, gene burden study, mutational signature study, and copy number variation study.

II. Literature Review

Recently, NLP methods for EHR-based computational phenotyping have seen extensive development beyond basic term and keyword extraction. One focus of recent studies is formulating
computational phenotyping as an unsupervised learning problem to automatically discover unknown phenotypes. The construction of richer features such as relations between medical concepts enables greater expressive power when encoding patient status, compared to terms and keywords. More advanced machine learning methods, such as deep learning, have also been increasingly adopted to learn the underlying patient representation. Computational phenotyping has facilitated biomedical and clinical research across many applications, including patient diagnosis categorization, novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction (DDI) and adverse drug event (ADE) detection, and downstream genomics studies.\textsuperscript{11-13}

Cancer is thought to occur as a consequence of the progressive accumulation of several somatic mutations. The history of this concept with a linear timeline is provided in a review by Alfred Knudson.\textsuperscript{14} All tissues accumulate DNA mutations over time. Most are repaired, many are inconsequential but a few may lead to or cause cancer. The few that lead to cancer almost certainly affect the processes and are referred to as driver mutations.\textsuperscript{15} The elegant, early epidemiology studies of Armitage and Doll estimated that, assuming the probability of mutation remains constant throughout life, 5-7 successive mutagenic events are required before cancer appears clinically.\textsuperscript{16} In a subsequent paper, they hypothesized that no more than two events were necessary, the first event resulting in a faster rate of multiplication that confers a selective advantage to the affected cells. The size of the mutant clone relative to other normal clones continuously increases, and the clinically apparent cancer manifests after the occurrence of a second mutation within the clone.\textsuperscript{17} Over 40 years later, Knudson pointed out that there is a range of numbers of genetic hits necessary to produce a malignancy from the single translocation to the double hit of retinoblastoma to the 4-5 hits, \textit{APC x 2, RAS, TP53 x2}, required for the development of colon carcinoma.\textsuperscript{14} Regarding
breast cancer, a recent study determined that approximately 20% of breast cancers contain a single
driver mutation, another 20% two driver mutations, a third 20% with three driver mutations and
15% contained four driver mutations.\textsuperscript{18} The development of massively parallel, high throughput
DNA sequencing has enabled the cataloging of somatic mutations in present in cancer. Mutations
are non-random and occur within sequence motifs. The motifs provide evidence from which to
infer the process that created the mutation. The groups that have put the greatest effort into this
are those headed by Michael Stratton\textsuperscript{3,19} and Serena Nik-Zaina\textsuperscript{5-7} at Cambridge. They have
produced, to date, 30 “Signatures of Mutational Processes in Human Cancer.” The underlying
process responsible for the mutations has been identified for some but not all of the Signatures.
The processes include, for example, aging, the activity of the AID/APOBEC family of cytidine
deaminases, pyrimidine dimer formation by ultraviolet light, exposure to tobacco mutagens and
aristolochic acid. Yet the underlying process responsible for the mutations has been identified for
some but not all of the signatures.

Previous pan-cancer studies have highlighted pathogenic germline variants for cancer
development. Gene ATM, BRCA1, BRCA2, BRIP1, PALB2, TP53, APC, BRCA2, NF1, PMS2,
and RB1 have been highlighted in The Cancer Genome Atlas (TCGA),\textsuperscript{20,21} the Collaborative
Oncological Gene-environment Study (COGS)\textsuperscript{22} and the Pediatric Cancer Genome Project
(PCGP).\textsuperscript{23} However, compared to the mounting amount of studies in somatic mutation, germline
variants were overlooked in previous years. Recently, germline variants have been identified as a
marker for pediatric cancer\textsuperscript{23}. Another study by Huang et al.\textsuperscript{24} suggested that germline variants
and somatic mutations affected the same residues for cancer development. On the other side, when
investigating variants, a frameshift mutation leads to an altered reading frame due to insertions or
deletions. Studying single nucleotide polymorphism (SNP) in a DNA sequence without
considering frameshift mutations, would potentially introduce bias because one might not be comparing the same SNPs in DNA sequence.

Recent advances in sequencing technology have overcome the problem of scale when studying DNA sequences. Abundant genome sequencing data are now available in different public repositories. The mounting scale of data also poses another challenge of computation ability. Conventional learning algorithms relies heavily on data representations, which are typically designed by domain expertise or human engineers. The human efforts make it difficult to scale the study to DNA sequencing studies, whereas deep learning can be unsupervised once set in motion. Deep learning has recently emerged based on big data, the power of parallel computing and sophisticated algorithms are exponentially more efficient than the conventional models in learning intricate patterns from high-dimensional raw data with little guidance.

Most cancers have long latency periods with the somatic mutations accumulating over time. Before there is any histologic evidence of cancer, histologically normal tissue and premalignant lesions have been demonstrated to contain molecular aberrations that are associated with malignancy. Histologically normal tissue from women at the normal risk for breast cancer show loss of heterozygosity or allelic imbalance of small segmental deletions at loci of potential tumor suppressor genes, DNA methylation of the promoters of tumor suppressor and other genes, telomere shortening and loss of TP53 function. In normal tissues at high risk for breast cancer, such as normal breast tissue adjacent to breast cancer or the contralateral breast, these changes persist, and are increased and accompanied by aneuploidy, increased genomic instability, a wide range of gene expression differences, development of large cancerized fields, and increased proliferation.
III. Innovation

We developed a model for computational phenotyping to generate breast cancer clinical phenotypes; developed a novel model to discover breast cancer mutational signatures and described the underlying etiology; developed a deep learning model for cancer classification; Furthermore, we designed a case-control study for breast cancer benign biopsy. These studies offered new insights in the below dimensions: 1) A system that combines structured EHR data and unstructured clinical narratives for breast cancer phenotyping. In addition to a well-tuned model, we also delivered a gold-standard data set with rich, validated information for further breast cancer research. 2) By studying germline variants, RNA level, methylation data, and clinical phenotypes, we uncovered the etiology associated with the signatures. 3) using the developed deep learning model, we were able to classify cancers using the unbiased gene reads that were constructed with germline variants information, somatic mutation information, SNP, INDEL, and motif information. Interpreting the results has uncovered new cancer genes. 4) Identification of biomarkers that would elevate cancer risk predication accuracy and inform the underlying mutational processes.
CHAPTER 2 COMPUTATIONAL PHENOTYPING

I. Computational Phenotyping

A phenotype is an expression of the characteristics that result from genotype variations and an organism’s interactions with its environment. A phenotype may consist of physical appearances (e.g., height, weight, BMI), biochemical processes, or behaviors. In the medical domain, phenotypes are often summarized by experts on the basis of clinical observations. Nationwide adoption of Electronic Health Records (EHRs) has given rise to a large amount of digital health data, which can be used for secondary analysis. Typical EHRs include structured data such as diagnosis codes, vitals and physiologic measurements, as well as unstructured clinical narratives such as progress notes and discharge summaries. Computational phenotyping aims to automatically mine or predict clinically significant, or scientifically meaningful, phenotypes from structured EHR data, unstructured clinical narratives, or their combination.

As summarized in a 2013 review by Shivade et al., early computational phenotyping studies were often formulated as supervised learning problems wherein a predefined phenotype is provided, and the task is to construct a patient cohort matching the definition’s criteria. Many of these studies relied heavily on structured and coded patient data; for example, using encodings such as International Classification of Disease, 9th Revision (ICD-9), its successor the 10th Revision (ICD-10), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), RxNorm, and Logical Observation Identifiers Names and Codes (LOINC). On the other hand, the use of natural language processing (NLP) for EHR-based computational phenotyping has been limited to term and keyword extraction.
Structured and unstructured EHR data are often stored in vendor applications or at a healthcare enterprise data warehouse. Typical EHR data are usually managed by a local institution’s technicians and are accessible to trained personnel or researchers. Institutional Review Boards at local institutions typically grant access to certain patient cohorts and certain parts of EHRs. Database queries can then be written and executed to retrieve desired structured and unstructured EHR data. In addition to hospital-collected data stored in EHRs, research data are increasingly available, including public databases such as PubMed, Textpresso, Human Protein Interaction Database (HPID), and MeInfoText. With growing amount of available data, efficient identification of relevant documents is essential to the research community. Information retrieval systems have been developed to identify text corresponding to certain topics or areas from EHR data across multiple fields. CoPub Mapper ranks co-occurrence associations between genes and biological terms from PubMed. iHOP links interacting proteins to their corresponding databases and uses co-occurrence information to build a graphical interaction network. We refer the reader to the following reviews for more details: a survey for biomedical text mining in cancer research, a survey for biomedical text mining, a survey for web mining.

II. Distant recurrence and natural language processing

Distant recurrences are defined as metastasis of the primary breast tumor to lymph nodes or organs beyond the loco-regional pathological field. Nodes located within the loco-regional field include ipsilateral axillary, ipsilateral internal mammary, supraclavicular, and intramammary lymph nodes. Distant lymph nodes beyond the loco-regional field include cervical, contralateral axillary, and contralateral internal mammary lymph nodes. The most common sites of metastasis to organs are the bone, brain, lung, and liver. It is important to distinguish between local and
distant recurrences for several reasons: the categorization informs treatment decision-making and directs studies analyzing outcomes of local versus distant recurrences. Most importantly, the 10-year survival rates are much lower for distant recurrences as compared to local recurrences (56% after an isolated local recurrence as opposed to 9% after distant metastasis). The delineation can be an important prognostic marker for mortality.

The emerging cancer prognosis research has directed efforts towards identifying distant cancer recurrence events accurately and efficiently. The National Program of Cancer Registries (NPCR) was launched to capture cancer patient information and one of its major tasks is to capture disease prognosis status for each cancer patient. However, many tumor registries fail to accurately identify cancer recurrences due to the significant human effort required for data maintenance. Manual chart review is one of the traditional methods used to identify breast cancer recurrences. Unfortunately, chart review is a time-consuming and costly process. It limits the number of samples available for research and is not feasible for large cohort studies. Furthermore, it is subject to human error in data analysis.

In this study, we aim to develop a model to identify distant recurrences within a cohort of breast cancer patients. To develop the model, we utilize data collected in Northwestern Medicine Enterprise Data Warehouse (NMEDW), which is a joint initiative across the Northwestern University Feinberg School of Medicine and Northwestern Memorial HealthCare. The NMEDW houses the EHR for about 6 million patients. Both structured and unstructured data are available in the NMEDW. Structured data typically capture patients’ demographic information, lab values, medications, diagnoses, and encounters. Although readily available and easily accessible, studies have concluded that structured data alone are not sufficient to accurately infer phenotypes. For example, ICD-9 codes are mainly recorded for administrative purposes and are influenced by
billing requirements and avoidance of liability.\textsuperscript{36,37} Consequently, these codes do not always accurately reflect a patient’s underlying physiology. Furthermore, not all patient information (such as clinicians’ observations and insights) is well documented in structured data.\textsuperscript{38} As a result, using structured data alone for phenotype identification often results in low performance.\textsuperscript{35} The limitations associated with structured data for computational phenotyping have encouraged the use of clinical narratives, which typically include clinicians’ notes, observations, referring letters, specialists’ reports, discharge summaries, and records of communication between doctors and patients.\textsuperscript{39} These clinical narratives contain rich descriptions of patients’ disease assessment, history, and treatments. However, the clinical narratives are not readily accessible without the use of natural language processing (NLP). The abundance of information in the free text makes NLP an indispensable tool for text-mining.\textsuperscript{55-57}

Our goal is to develop such a system that combines structured EHR data and unstructured clinical narratives to accurately and efficiently identify distant recurrences in breast cancer. Such a model can be easily replicated and requires a minimum amount of human effort and input.

\textbf{III. Related Work}

Different NLP applications have been developed to identify breast cancer recurrences. Carrell et al.\textsuperscript{58} proposed a method to identify breast cancer sub-cohorts with ipsilateral, regional, and metastatic events using the concepts identified within the free text. The binary classification model achieved an F-measure scores of 0.84 and 0.82 in the training set and test set, respectively. However, the model could not distinguish a local recurrence from a distant recurrence. In addition, defining the number of hits in the system to segment the documents required substantial effort. Using morphology codes and anatomical sites from pathology reports, Strauss et al.\textsuperscript{12} were able to
identify recurrences. However, their approach required that the pathology reports be well-documented under a standard format. However, the majority of distant recurrences in breast cancer have been diagnosed clinically rather than pathologically.\textsuperscript{59} It has been challenging to identify distant recurrences from pathology reports because they are not usually recorded as clinical diagnoses in the reports. Haque et al.\textsuperscript{13} applied a hybrid approach to identify breast cancer recurrences using a combination of pathology reports and EHR data. They achieved a relatively high NPV of 0.995 and a relatively low PPV of 0.65. This model also required a minimum amount of ten percent manual chart review, which is still fairly time-consuming. In addition, the model was not able to distinguish between local, regional, or distant recurrences. NLP has also been applied to attempt retrieving distant recurrences for other types of cancer. Lauren et al.\textsuperscript{60} tried to identify distant recurrences in prostate cancer from clinical notes, radiology reports, and pathology reports. They concluded that NLP could be used to identify metastatic prostate events more accurately than claim data.

Clinical narratives are known to have high-dimensional feature spaces, few irrelevant features, and sparse instance vectors.\textsuperscript{61} These problems were found to be well-addressed by SVMs,\textsuperscript{61} which also have been recognized for their generalizability and are widely used for computational phenotyping.\textsuperscript{55,62–67} Carroll et al.\textsuperscript{66} implemented a SVM model for rheumatoid arthritis identification using a set of features from clinical narratives using the Knowledge Map Concept Identifier (KMCI).\textsuperscript{68} They demonstrated that a SVM algorithm trained on these features outperformed a deterministic algorithm.

A combination of structured data and narratives for phenotyping have been found to improve model performances. DeLisle et al.\textsuperscript{69} implemented a model to identify acute respiratory infections. They used structured data combined with narrative reports and demonstrated that the inclusion of
free text improved the PPV score by 0.2–0.7 while retaining sensitivities around 0.58–0.75. In a study of the identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis, Lin et al.\textsuperscript{70} obtained an F-measure of 0.83 in a performance evaluation. Liao et al.\textsuperscript{71} implemented a penalized logistic regression as a classification algorithm to predict patients’ probabilities of having Crohn’s disease and achieved a PPV score of 0.98. Both Lin’s and Liao’s methods experimented with a combination of features from structured EHR and NLP-processed features from clinical narratives. Their studies showed that the inclusion of NLP methods resulted in significantly improved performance.

IV. Methodology

A. Cohort Description

Patients diagnosed with breast cancer between 01/01/2001 and 12/31/2015 were drawn from NMEDW. Patients were identified by ICD-9 codes. In total, 19,874 females were included. Within this cohort, only cases with at least one surgical pathology report documented in the desired time window were selected. In total, 6,899 subjects were identified and included in this study. The workflow to generate this data set is presented in Figure 1.

To establish a gold standard for algorithm development, each patient was assigned a definite distant recurrence status (‘Yes’ or ‘No’) according to manual chart review. In total, 1,995 subjects
were annotated twice by two annotators and were included for model training and validation. The inter-rater agreements for the two annotators were measured by Cohen’s kappa score, and the obtained score is 0.87.72 The items without agreements were resolved by a discussion between the two annotators. The other 4,904 subjects were annotated once by annotators and were used as an independent set for model generalization test. These annotations were conducted over a span of 15 months (completed September 2017).

The 1,995 double-annotated subjects were randomly split into a cross-validation set and a held-out test set according to a 7:3 ratio. In the cross-validation set, five-fold cross-validation was applied with the 1,396 samples. Among these 1,396 samples, 138 distant recurrence events were identified; among the 599 samples in held-out test set, 55 distant recurrences were identified. In the generalization test set, 443 distant recurrences were identified among the 4,904 samples. The cohort distribution is shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Distant Recurrence</th>
<th>Percentage (%)</th>
<th>Overall percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Double-annotated set</strong></td>
<td>1,995</td>
<td>193</td>
<td>9.87%</td>
<td></td>
</tr>
<tr>
<td>Cross-validation set</td>
<td>1,396</td>
<td>138</td>
<td>9.89%</td>
<td>9.22%</td>
</tr>
<tr>
<td>Held-out test set</td>
<td>599</td>
<td>55</td>
<td>9.19%</td>
<td></td>
</tr>
<tr>
<td><strong>Single-annotated set</strong></td>
<td>4,904</td>
<td>443</td>
<td>9.03%</td>
<td></td>
</tr>
</tbody>
</table>

**B. Structured Clinical Data**

Automated SQL codes were developed to query structured data from NMEDW and from cancer registry. In total, 18 structured clinical variables were retrieved or derived. The variable names and corresponding categories or values are displayed in Table 1. Demographic data such as the age of diagnosis, race, smoking history, alcohol usage, family cancer history, and insurance type were
queried. Smoking history was categorized as ‘Yes’, ‘No’, ‘Ex-smoker’, or ‘Unknown’. Alcohol usage was categorized as ‘No’, ‘Moderate’, 'Heavy’, 'Former’, or ‘Unknown’. Tumor characteristics and biomarkers, such as estrogen receptor (ER), progesterone receptor (PR), HER2, P53, nodal positivity, histology, tumor grade, and tumor size are retrieved. Nodal positivity was categorized as ‘Positive’, ‘Negative’, or ‘Unknown’. The variable histology and nodal positivity were selected, because subjects with invasive ductal breast cancer or positive lymph nodes were more likely to develop a distant recurrence compared to those that have ductal in situ or negative lymph nodes. Primary surgery type is categorized as ‘Breast conservation surgery’, ‘Mastectomy’, ‘No’, or ‘Unknown’.

Table 2: The name and corresponding categories (values) of the 18 retrieved structured clinical variables. IDC is invasive ductal carcinoma, DCIS is ductal carcinoma in situ, ILC is invasive lobular carcinoma, Network Category is the network of patient’s insurance plan.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>age of diagnosis</td>
<td>Continuous</td>
</tr>
<tr>
<td>Race</td>
<td>White, Black, Asian, Other</td>
</tr>
<tr>
<td>smoking history</td>
<td>Yes, No, Ex-smoker, Unknown</td>
</tr>
<tr>
<td>alcohol usage</td>
<td>No, Moderate, Heavy, Former, Unknown</td>
</tr>
<tr>
<td>family cancer history</td>
<td>Yes, No, Unknown</td>
</tr>
<tr>
<td>insurance type</td>
<td>Network Category</td>
</tr>
<tr>
<td>estrogen receptor</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>progesterone receptor</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>HER2</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>P53</td>
<td>Positive, Negative, Unknown</td>
</tr>
<tr>
<td>nodal positivity</td>
<td>Positive, Negative, or Unknown</td>
</tr>
<tr>
<td>Histology</td>
<td>IDC, DCIS, ILC, Unknown</td>
</tr>
<tr>
<td>grade</td>
<td>Grade1, Grade2, Grade3, Unknown</td>
</tr>
<tr>
<td>Size</td>
<td>0-2cm, 2cm-5cm, &gt;5cm, Unknown</td>
</tr>
<tr>
<td>surgery type</td>
<td>Mastectomy, Breast conservation surgery, Unknown</td>
</tr>
<tr>
<td>Deceased</td>
<td>Yes, No</td>
</tr>
<tr>
<td>targeted therapy</td>
<td>Yes, No</td>
</tr>
<tr>
<td>radiation</td>
<td>Yes, No</td>
</tr>
</tbody>
</table>

Additional clinical variables were derived to help identify distant recurrences. Variables of deceased, targeted therapy, and radiation were developed. The deceased variable was a binary
variable to indicate whether a patient deceased before the age of 75. Intuitively, patients with
distant recurrences might have a shorter survival length compared to the women who do not have
distant recurrences. After a discussion with a domain expert (co-author SK), we chose the age of
75 as the cutoff. Another variable ‘targeted therapy’ is a binary variable created to indicate whether
the patient had taken any of the following drugs: ‘Afinitor’, ‘Everolimus’, ‘Bevacizumab’,
‘Avastin’, ‘Ibrance’, or ‘Palbociclib’. These drugs were prescriptions for patients with distant
recurrences. An additional variable radiation was a binary variable indicating whether the subject
had received radiation treatment at the site of metastases, such as brain, lung, or bone. This variable
was derived from the intuition that patients receiving radiation at a site different from the primary
tumor were at a higher chance of having distant recurrences.

C. Clinical Narratives

We queried the NMEDW for clinical narratives generated before May 2016 (the start time of
manual chart review) or the date when the patient was censored. All inpatient and outpatient notes
were retrieved without any provider type restriction. The retrieved clinical narratives include
progress notes, pathology reports, telephone encounter notes, assessment and plan notes, problem
overview notes, treatment summary notes, radiology notes, lab notes, procedural notes, and
nursing notes. Only notes generated after the diagnosis of breast cancer were retrieved. We only
included the notes having at least one mention of ‘breast’. After retrieving the narratives, we first
preprocessed the corpus by removing duplicate copies, tokenizing sentences, and removing non-
English symbols. Following these preprocessing steps, we annotated the medical concepts in the
sentences using MetaMap, an NLP application to map the biomedical text to the UMLS
Metathesaurus. The surrounding semantic context was determined. CUIs that were tagged as
negated by NegEx\textsuperscript{75} were excluded (NegEx is a negation tool configured in MetaMap). If multiple CUIs were mapped, the one with maximum MMI score (a score ranked by MetaMap) was retained. In order to completely and accurately exclude negations or unrelated contextual cues, such as a differential diagnosis event, sentences with negative contextual features (e.g., ‘no’, ‘rule out’, ‘deny’, ‘unremarkable’) and uncertain contextual features (e.g., ‘risk’, ‘concern’, ‘worry’, ‘evaluation’) were also removed. This customized list of contextual features was obtained from the development corpus.

**D. Feature Generation**

To focus our NLP efforts, we identified a set of target distant recurrence concepts with the help of sample notes. We reviewed a development corpus of ten randomly selected samples’ notes with distant recurrences and extracted partial sentences that were related to a breast cancer distant recurrence. The initial set contains 20 partial sentences. These partial sentences were tagged by MetaMap, and the CUIs corresponding to each concept was obtained. The customized dictionary contains 83 CUIs. After data preprocessing and concept mapping, only CUIs with highest MMI score that also fell within the customized dictionary were used as features for model training. CUIs with MMI score smaller than one were filtered and excluded. Following this feature selection, there were 83 narrative-based features remaining for inclusion in the machine learning algorithm. In addition to the obtained CUI features, the 18 structured clinical variables described above were used as additional features.

**E. Prediction Model and Evaluation**

We used support vector machine (SVM) to develop an algorithm to predict whether patients had distant recurrences. SVMs have been widely used for computational phenotyping.\textsuperscript{55,62-67} We
applied linear kernel type for the SVM models. In our experiments, we trained four baseline classifiers on different feature types: a full set of medical concepts tagged by MetaMap, a filtered set of medical concepts tagged by MetaMap, only the structured clinical data, and a standard bag of words from clinical narratives. TfIDFVectorizer class in scikit-learn was used to convert the raw documents to a matrix of TF-IDF features to assemble a bag of words. In the full MetaMap and bag of words, Chi-square test was applied to select features before training the model to remove the common words that exist in clinical narratives. Only top 5% features were retained for modeling.

In the model evaluation, we chose area under curve (AUC) score as a measurement metric because this is a skewed cohort with low event rate. The output of our SVM model is probabilities, though in practice, various thresholds result in different true positive/false positive rates and AUC score considers all possible thresholds. To better demonstrate the thresholds and model performance, the corresponding receiver operating characteristic (ROC) curves for the different methods were evaluated. Cross validation performance depends on the randomly shuffled split of the training dataset into multiple folds. In order to obtain robust performance statistics, each five-fold cross validation was replicated 20 times using shuffled stratified splits initialized with different random seeds.

V. Experiment results

As demonstrated in Table 3, clinical data with a significant difference between the recurrence group and the non-recurrence group in the double-annotated training set are presented. Compared to the non-recurrence patients, women with recurrences had a higher percentage of nodal positivity and higher grade of tumor, were more likely to be diagnosed with invasive ductal carcinomas, had
more radiation performed at the metastasis site, had received more targeted therapies, and were more likely to die before the age of 75.

Using SVM as a prediction model, the AUC scores obtained from the cross-validation were reported in Table 4. To note, the model applied was an SVM model with linear kernel (C equaled 1, and gamma was set as ‘auto’ in the python package ‘sklearn.svm’).

Table 3: Descriptive summaries of 1,995 subjects’ clinical data. The significance test is performed between the recurrence group and the non-recurrence group. Only data with P-values less than 0.05 are presented. DR stands for distant recurrence. The mean and standard deviation are calculated for continuous variables. Numbers and percentages are presented for categorical variables. P-values are obtained using Student’s t-test for continuous variables and Chi-squared test for categorical variables.

<table>
<thead>
<tr>
<th></th>
<th>Double-annotated set N=1,995</th>
<th>DR N=193</th>
<th>No DR N=1,802</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodal positivity (%)</td>
<td>544 (27.3%)</td>
<td>103 (53.4%)</td>
<td>441 (24.5%)</td>
<td>1.4e-14</td>
</tr>
<tr>
<td>Histology (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDC</td>
<td>1,530 (76.7%)</td>
<td>174</td>
<td>1,356 (75.2%)</td>
<td></td>
</tr>
<tr>
<td>DCIS</td>
<td>279 (14.0%)</td>
<td>3 (1.6%)</td>
<td>276 (15.3%)</td>
<td></td>
</tr>
<tr>
<td>ILC</td>
<td>155 (7.8%)</td>
<td>15 (7.8%)</td>
<td>140 (7.8%)</td>
<td></td>
</tr>
<tr>
<td>Grade (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1</td>
<td>458 (23.0%)</td>
<td>16 (8.3%)</td>
<td>442 (24.5%)</td>
<td></td>
</tr>
<tr>
<td>Grade 2</td>
<td>851 (42.7%)</td>
<td>73 (37.8%)</td>
<td>778 (43.2%)</td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>665 (33.3%)</td>
<td>101</td>
<td>564 (31.3%)</td>
<td></td>
</tr>
<tr>
<td>Deceased (%)</td>
<td>157 (7.9%)</td>
<td>98 (50.8%)</td>
<td>59 (3.3%)</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Radiation (%)</td>
<td>67 (3.4%)</td>
<td>52 (26.9%)</td>
<td>15 (0.8%)</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Targeted therapy (%)</td>
<td>60 (3.0%)</td>
<td>44 (22.8%)</td>
<td>16 (0.9%)</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>
The AUC score obtained in our proposed model was 0.92 (SD=0.01). The performance of our proposed model significantly outperformed the other four baselines. The P-value for Student’s t-test was 0.0004 comparing our proposed model with the second-ranked model of Filtered MetaMap.

To illustrate, the corresponding receiver operating characteristic (ROC) curves for the different methods are plotted in Figure 2.

We trained an SVM model on the training set (1,396 samples) and then predicted labels on the held-out test set (599 samples). Comparing the predicted probabilities and the annotated labels, the obtained AUC scores are presented in Table 5. The AUC score obtained in our proposed model was 0.95. The model with NLP-features, Filtered MetaMap also had a notable performance of 0.93. The performance in our proposed model again outperformed all the baseline models.

In addition to our training and validation analyses, we applied our fitted model to predict labels on the generalization set, which contained 4,904 single-annotated samples. In this generalization test, we obtained an AUC score of 0.92, which had a similar performance as the held-out test.
From the fitted SVM model using the 1,396 samples in the training set, we retrieved the coefficient scores for each feature. The top 15 ranked coefficient scores and their corresponding variable names appear in Table 6. Three of the clinical variables (radiation, deceased, and targeted therapy) were highly ranked on the list. These three variables were treatment or outcome variables. The rest of the top-ranked features were concepts obtained from clinical narratives. Most of the CUIs were either related to metastases events or related to the metastatic sites that breast cancer could spread to. The term ‘IXEMPRA’ is a prescription medicine used for locally advanced breast cancer or breast cancer with distant recurrences.
Table 6: Top 15 variables with their corresponding coefficients

<table>
<thead>
<tr>
<th>CUI</th>
<th>Name</th>
<th>Coefficient</th>
<th>Partial Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0153678</td>
<td>Secondary malignant neoplasm of pleura</td>
<td>1.00</td>
<td>cancer metastatic to pleura</td>
</tr>
<tr>
<td>Radiation</td>
<td>Clinical Variable</td>
<td>0.90</td>
<td>metastatic cancer to pleura</td>
</tr>
<tr>
<td>Deceased</td>
<td>Clinical Variable</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Targeted therapy</td>
<td>Clinical Variable</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>C0153690</td>
<td>Secondary malignant neoplasm of bone</td>
<td>0.76</td>
<td>metastases to bone, bone metastases</td>
</tr>
<tr>
<td>C1967552</td>
<td>IXEMPRA</td>
<td>0.71</td>
<td>ixempra</td>
</tr>
<tr>
<td>C0278488</td>
<td>Carcinoma breast stage IV</td>
<td>0.70</td>
<td>metastatic breast cancer, breast cancer stage iv, metastatic breast</td>
</tr>
<tr>
<td>C0494165</td>
<td>Secondary malignant neoplasm of liver</td>
<td>0.62</td>
<td>liver metastases, liver metastatic disease, metastatic disease liver, metastases to the liver, liver</td>
</tr>
<tr>
<td>C0220650</td>
<td>Metastatic malignant neoplasm to brain</td>
<td>0.59</td>
<td>brain metastases</td>
</tr>
<tr>
<td>C1266909</td>
<td>Entire bony skeleton</td>
<td>0.39</td>
<td>bone</td>
</tr>
<tr>
<td>C2939420</td>
<td>Metastatic Neoplasm</td>
<td>0.27</td>
<td>metastatic disease</td>
</tr>
<tr>
<td>C0036527</td>
<td>Metastatic to</td>
<td>0.25</td>
<td>metastatic</td>
</tr>
<tr>
<td>C0027627</td>
<td>Neoplasm Metastasis</td>
<td>0.25</td>
<td>metastatic disease</td>
</tr>
<tr>
<td>C0346993</td>
<td>Secondary malignant neoplasm of female</td>
<td>0.23</td>
<td>metastatic breast cancer to the</td>
</tr>
<tr>
<td>C1522484</td>
<td>Metastatic qualifier</td>
<td>0.22</td>
<td>metastatic</td>
</tr>
</tbody>
</table>

VI. Discussion

In this study, we combined 83 features from unstructured clinical narratives and 18 features from structured clinical data to identify distant recurrences in breast cancer. Clinical narratives were extracted from progress notes, pathology reports, telephone encounter notes, assessment and plan notes, problem overview notes, treatment summary notes, radiology notes, lab notes, procedural notes, and nursing notes generated after diagnosis of primary breast cancer. The clinical narratives were tagged by NLP application MetaMap to generate UMLS concepts. After filtering out concepts that were not in the customized dictionary, the remaining concepts were combined with the structured clinical data to train an SVM model for distant recurrence identification. We were
able to identify structured clinical variables that could stratify the groups of women with and without distant recurrences. Using such a method, we obtained an AUC score of 0.95 and 0.92 in our external held-out test and generalization test.

During the feature coefficient study, we found that the features “secondary malignant neoplasm of pleura, radiation, deceased, targeted therapy, and secondary malignant neoplasm of bone” were the top-ranked features. Intuitively, women with distant recurrences have a higher chance of receiving radiation at the metastatic site and of receiving targeted therapy compared to those without distant recurrences. They are also more likely to have a lower survival rate. The most common sites of metastasis to organs were the bone, brain, lung, and liver. In our study, we found the mentions of metastatic to bone, liver, and brain were also top-ranked. The terms ‘metastatic’ and ‘breast cancer’ were also more likely to appear in the clinical notes of patients with distant recurrences.

Progress notes are notably telegraphic. Also, many excessively busy residents and senior clinicians create notes by simply copying and pasting previous encounter notes, while making only minor updates for the most recent appointment. This results in many notes that differ in ‘critical’ content still scoring highly on the overall measures of similarity. The same applies to the full set of MetaMap concepts, which is similar to the bag of words. To include only highly associated features in our model, we removed the common concepts or words in the notes. Chi-square test was applied to select features before training the model. Only the top 5% features were retained for full MetaMap and bag of words modeling. This test might have the potential for overfitting in cross-validations. Indeed, we saw a lower performance in the held-out test for full MetaMap and bag of words. To adjust this problem, we tested different thresholds for the Chi-square test selection. However, we found 5% ended with the best results.
Identifying breast cancer distant recurrence in clinical data sets is important for clinical research and practice. Annotation of distant recurrence is difficult using standard EHR phenotyping approaches and are commonly beyond the scope of manual annotation efforts by cancer registries.

A model using natural language processing, EHR data, and machine learning to identify distant recurrences in breast cancer patients allows more accurate data-mining and significantly less time-consuming manual chart review. We expect that by minimally adapting the positive concept set, this study has the potential to be replicated at other institutions with a moderately sized training dataset. In this study, we generated features using sentences extracted from the clinical narratives combined with structured data. The training and testing data sets were cross-annotated in the process, which offered a solid ground truth for the study. Replicating this model requires minimal outside effort. We offered the customized dictionary in this study, so a user can retrieve the required notes and clinical structured data in order to replicate this study. After the rigorous manual chart review and feature retrieval, our data set has offered a gold-standard data set with rich, validated information for further breast cancer research.

When replicating this study at another institution, there is a chance that one will not be able to find the structured clinical data in their databases. If this is the case, some of the structured data can be found from other resources. Variables of ‘histology’ and ‘lymph node status’ can be extracted from pathology reports using a rule-based system. For example, expressions of ‘total lymph nodes’, ‘total lymph nodes number positive’, ‘axillary lymph nodes examined’, ‘axillary lymph nodes examined number of positive versus total’ can be used to extract lymph node status from pathology report at our institution. Survival information can be found in the administrative billing system.
VII. Conclusions of computational phenotyping

The NLP pipeline cannot characterize the context of features. Clinical narratives contain patients’ concerns, clinicians’ assumptions, and patients’ past medical histories. Clinicians also record diagnoses that are ruled out or symptoms that patients denied. Our next aim will be that such conditions, mentions, and feature relations will be extracted to better distinguish differential diagnoses. Generalized relation and event extraction, rather than binary relation classification, will be conducted. To this end, graph methods are a promising class of algorithms and should be actively investigated.76,77

In the future, we will test our data with different machine learning models. In this study, we have chosen SVM model with linear kernel for interpretation purposes. Other models might result in better performance.

We will also aim to address the heterogeneity problem in clinical narratives. It is a common problem in clinical narratives due to the variance in physicians’ expertise and behaviors.78 Features derived from clinical narratives included in this study were extracted from notes generated by clinicians with different specialties and professional levels of expertise. As a result, some content was not relevant to the breast cancer distant recurrence event, even though we had limited the notes to include the mention of ‘breast’. For example, a liver cancer metastasis to the breast from a primary liver tumor would be difficult to identify. We will need to resolve the heterogeneity in clinical narratives.

We developed a machine learning model by combining structured clinical data and unstructured clinical narratives in order to identify distant recurrence events in breast cancer. We demonstrated the high accuracy and efficiency of our model, using cross-validation, held-out test evaluation, and a further generalization set evaluation. Our proposed model allows for more accurate and efficient
identification of distant recurrences than single modality models using either clinical narratives or structured clinical data. Thus, our model is a significantly less time-consuming and practical alternative to manual chart review. This is particularly relevant in an era when evidence-based medicine receives growing attention and there is more emphasis on computational phenotyping and data-driven discovery. This model would also be valuable and applicable to research in other medical fields beyond breast cancer.
CHAPTER 3 Mutational Processes in Breast Cancer

I. Introduction

Breast cancer is a heterogeneous disease with major subtypes related to estrogen receptor (ER), progesterone receptor (PR) and HER2 expression, while ever-finer subtypes continue to be defined (1,2). Continuing progress in subtype-specific breast cancer prevention is most likely to require an understanding of the etiology of the different subtypes. The development of massively parallel, high throughput DNA sequencing has enabled the identification and cataloging of somatic mutations in cancer. It has been reported that tumors of the HER2 subtype are enriched with mutations hypothesized to result from the action of APOBEC enzymes (3). The wide array of established endocrine risk factors for breast cancer are well described. However, no DNA mutations have been associated with exposure to the ovarian steroids, which may not have been considered surprising as estrogen, for example, does not induce mutations in standard test systems (4).

Mutations are non-random and occur within sequence motifs. These motifs provide evidence from which we can infer the process that created the mutation. Recent studies led by investigators at the Welcome Trust Sanger Institute (WTSI) presented the somatic mutation data as a 96-element vector, which captures the immediate 5’ and 3’ neighbors of the mutated nucleotides. Employing non-negative matrix factorization (NMF) to derive signatures, at the end of 2018 the number of “mutational signatures” produced by these groups stands at 30 (5,6). While these signatures provide important insights into the processes responsible for somatic mutation, in vitro validation is incomplete, and the etiology of several signatures remains unknown.

In the various somatic mutation signature studies, sites of nucleotide change were investigated in the center of trinucleotide motifs (6,7), in the center of pentanucleotide (8), or in a sequence-independent context (3,9,10). Recently, the gene coding direction has been described to be informative for mutational profiles (11). Therefore, in this study, we have developed a new method to create mutational signatures, which
incorporates the specific nucleotide change, the immediate 5’ and 3’ neighbors of the mutated nucleotide; adds the direction of transcription and excludes hypermutators. Using this method and data from 1,000 breast cancer specimens in The Cancer Genome Atlas (TCGA), we have identified four signatures, one of which is completely novel and unique to breast cancer. Additionally, one of these, which we have designated Process Signature Oxidative 2 (POx2), is enriched for triple negative tumors and young patients.

II. Methods

A. Somatic mutations

Whole exome sequencing data were retrieved from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp). The VCF files containing somatic mutations were annotated with ANNOVAR (51) and SNPEFF(52). Mutational processes have been reported within hypermutated samples (6). In this study, we removed 44 samples out of 1,044 with mutation counts larger than the 98th percentile or lower than the second percentile among all samples to reduce bias caused by outliers. To note, only single-nucleotide variants were included for signature analysis. In total, 72,378 somatic mutations from 1,000 samples were retained, with an average of 72 (95% CI : 68, 76) somatic mutations.

B. Signature modeling

We developed a method in which the specific nucleotide change (e.g., C>T), the nucleotides immediately on the 5’ and 3’ sides of the mutation (e.g., 5’: C; 3’: G) are considered. In addition, the transcription directions of the genes were accounted to reflect the flanking nucleotide at 3’ and 5’ using the database of RefSeq (53). The summary of these mutation characteristics forms a mutational profile for each tissue sample. Putting multiple samples’ profiles together form a sparse matrix with the number of samples as rows (1,000) and the mutation characteristics as columns (192). Nonsmooth nonnegative matrix factorization (54) (nsNMF) was applied to enable the discovery of intrinsic patterns in this sparse matrix. The first value where the Residual Sum of Squares (RSS) curve presents an inflection point was used to determine the number of signatures (55). In total, four signatures were discovered among the 1,000 tissue
samples.
The outputs of nsNMF consist of an H matrix and a W matrix. The matrix H (dimension of 4 × 192) was used to infer mutational processes. The numbers in matrix W (dimension of 1,000 × 4) correspond to each samples’ signature exposure levels. This matrix was interpreted as each tissue sample’s accumulated exposure effect to the mutational burden. To explore whether the signatures were active in any other cancers, the signature study was repeated in other 17 major cancers that have more than 300 samples from TCGA, including Bladder Urothelial Carcinoma (BLCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Brain Lower Grade Glioma (LGG), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Prostate adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Thyroid carcinoma (THCA), and Uterine Corpus Endometrial Carcinoma (UCEC). The similarity between signatures was measured by cosine similarity score. We also mimicked our signatures to the WSTI signature patterns for similarity exploration.

C. Gene expression level change

Gene counts derived from RNA-Seq were assessed and quantified with HTseq(56). In the 1,000 breast cancer samples, 993 have RNA-Seq data and were retrieved. EdgeR from the Bioconductor package was used to produce normalized counts per million (CPM) for each gene (57). Pearson’s product-moment correlation tests were performed between the signature exposure levels and the gene expression levels. Correlation coefficients and p-values were derived for each gene. Benjamini-Hochberg method was applied to convert the two-sided p-values to False Discover Rate (FDR) for multi-comparison correction. Gene ontology (GO) enrichment analysis was performed using Metascape (http://metascape.org) to reveal significant pathways (adjusted p < 0.05) enriched with the co-expressed genes.
D. Germline variants and somatic mutations discovery

In the 1,000 breast cancer samples, 992 have matched normal DNA. To study the associations between signature exposure level and germline variants, re-aligned BAM files sequenced from 992 matched blood samples or adjacent normal tissues were retrieved. The BAM files were processed following GATK’s best practices (58), including recalibrating, variant calling (HaplotypeCaller), and variant filtering. The produced VCF files containing germline variants were further annotated with ANNOVAR (51) and SNPEFF (52). To remove common variants, considering that we have 160 (16%) basal subtypes, any variant appeared in more than 20% of the cohorts was removed. SKAT (59) and gene burden analyses (60) were both applied to study the associations between genes and signature exposure levels. SKAT (59) is a kernel matrix-based model for rare variant association discovery. The input for SKAT is a matrix with rows as sample tissues and columns as genomic locations. The output for SKAT is the signature exposure level. Gene burden analyses (60) were also performed. In short, variants were accumulated in each gene. The associations between the summed numbers and the signature exposure levels were evaluated. The association effects and the corresponding P-values were derived. FDR was calculated for multi-comparison correction. Genes were considered to be significantly associated with signature exposure levels if any of the tests was significant (FDR < 0.05).

We have also performed association study between the somatic mutations and the signature exposure levels. SKAT and gene burden tests were performed for the study. After examining the associations, genes were sorted by the derived FDR. To further select the genes that were most strongly associated with the gene expression levels, lasso regressions were fit using the signature exposure levels as outputs and the somatic mutation burden in each gene as inputs. The genes with FDR less than 5% were used as inputs and the signature exposure levels as outputs. Lambda=0.0005 in lasso regression was selected by 10-fold cross-validation.
E. Methylation discovery

Samples with methylation data measured by HumanMethylation450 (HM450) were retrieved. The HM450 targets 482,421 CpG sites and covers 99% of the RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5’UTR, first exon, gene body, and 3’UTR. The methylation $\beta$ values, which were used to measure the methylation levels of each CpG site, were retrieved. Within the breast cancer data set, 711 sets of methylation data were identified. Pearson’s product-moment correlation tests were performed between the $\beta$ values and the signature exposure levels. Probes with FDR less than 0.05 were considered to be significantly associated with the relevant signature exposure levels.

F. Interaction between gene expression level and signature exposure level

We evaluated the association between the signature exposure level and mortality with cox proportional hazard regressions, adjusting for age and stage. We then evaluated the interaction effects between the signature exposure levels and the gene expression levels on mortality. In statistics, an interaction arises when the effect of one predictor on the outcome depends on another predictor. Interaction terms between the signature exposure levels and the gene expression levels were included in the regression models. The interaction terms were evaluated by the Wald test z score, which is the coefficient divided by its standard error. The significant (FDR <0.05) genes were sorted by the FDR.

We performed stratified analyses if effect modification was suggested by significant interaction terms. We first used Kaplan-Meier survival curves to describe the association between the signature exposure level and mortality, stratifying on the gene expression level. The signature exposure level and gene expression level are continuous numbers. To generate the curves, we dichotomized the continuous numbers to binary variables. For both variables, numbers above the average were grouped into the high-level group, and others were grouped into the low-level group. Furthermore, we ran age and stage adjusted cox proportional hazard regression models to examine the association between the signature exposure level and mortality, stratifying on gene expression level.
III. Results

A. Mutational patterns

Using 72,378 somatic mutations identified in 1,000 breast cancer tissues from The Cancer Genome Atlas (TCGA), we have identified four mutational signatures, which we have named by the putative process that created them. Process Signature A (PSA) [A for APOBEC] correlates with the well-defined APOBEC signatures (WTSI Signatures 2; Figure 3; cophen similarity score: 88%) and Process Signature H (PSH) [H for hydrolysis] with the “aging” signature (WTSI Signature 1b; Figure 3; cophen similarity score: 96%), which is the putative result of the hydrolysis 5-methylcytosine. Process Signature Ox1 (PSOx1) [Ox for oxidative] and Process Signature Ox2 (PSOx2) are potentially new signatures (Figure 4). PSOx1 is mainly enriched with T-to-G (T>G) transversion mutations in the context GpTpNp>GpGpNp [N=A,C,G,T] and its complement NpApCp>NpCpCp. PSOx2 is dominated by G>T (C>A), G>C (C>G) and G>A (C>T) mutations.

B. Signatures A and H

Signatures A and H reprise 3 WTSI signatures. Their presence among our signatures gives us confidence that our method of signature development is commensurate with what has already been published. As these signatures have been well characterized, we will not discuss them further.
C. Process Signature Ox1

Although T>G mutations can occur consequent to template dislocation errors by DNA polymerase β, the majority of the T>G mutations in Process Signature Ox1 are most likely due to the insertion of an oxidized form of dGTP across from dA during polymerization that is not subsequently repaired. PSOx1 is unique to breast cancer, which behooves us to develop a plausible hypothesis to explain its generation exclusively in this organ. Examination of oxidants provides clues. Ionizing radiation produces a high proportion of T>G (A>C) transversions, which are distinct from the C>T (G>A) transitions induced by ultraviolet radiation. The GpTpNp>GpGpNp mutation motif forms a large part of the Ionizing Radiation b (IRb) mutational signature, which has been observed in the irradiated tumor cells of humans and mice. Like the IRb Signature, the T>G mutations observed in PSOx1 are characterized by a 5’G; the GpTpGp motif evidences the greatest number of mutations. γ-irradiation creates DNA damage via its
generation of hydroxyl radicals. Another source of hydroxyl radicals is that produced from hydrogen peroxide (H₂O₂) via the Fenton reaction, which is catalyzed by Fe(II) or Cu(I). A source of H₂O₂ unique to the breast derives from the metabolism of estrogen. The estrogens 17 β-estradiol (E₂) and estrone (E₁) are metabolized by two major pathways; the products of one of these pathways are the catechol estrogens 2-hydroxyestradiol (2OHE2) and 4-hydroxyestradiol (4OHE2). H₂O₂ can be generated from endogenous catechol estrogen-stimulated redox cycling. Fussell et al. have suggested that quantities of catechol estrogen metabolites in breast epithelial cells are in the submicromolar range which would be sufficient to generate H₂O₂. These catechol estrogens are localized to the endoplasmic reticulum and H₂O₂ can diffuse up to 50% of the diameter of the cell potentially placing it within the nucleotide pool. Hydroxyl radicals react where they are generated, which means that the metals for the Fenton reaction, most likely unincorporated intracellular ferrous iron, must be present in the pool. A>G mutations are also a component of PSOx1 in the motif 5’GAPurine3’. These are likely due to oxidation of dATP within the nucleotide pool. There are no complementary T>C mutations observed. This specific strand bias was also observed by Alexandrov et al. and attributed to transcription-coupled nucleotide excision repair (NER). C>A mutations in the ApCpAp motif are an additional feature of this signature. In sum, we hypothesize that mutations observed in Process Signature Ox1 result from hydroxyl radical initiated oxidation of dGTP and dATP within the nucleotide pool.
Figure 4: Mutational patterns of Process Signature Ox1 and Process Signature Ox2. The specific nucleotide change is outside the circles. The 5’ neighbors of the mutated nucleotide are at the inner circles. The 3’ neighbors of the mutated nucleotide are at the far end of bars. The green circles and pink circles are drawn at 30% and 60% of the y axis.

**D. Correlation of Process Signature Ox1 with RNA expression**

RNA expression has been correlated with PSOx1 and the individual RNAs with the most significant correlation are presented in Figure 5A. Many of the RNAs most significantly correlated are expressed mitochondrial genes, antisense transcripts and small nucleolar RNAs, in particular, small Cajal body-specific RNAs (SCARNAs). A common theme shared among the RNAs correlated with PSOx1 is that they are expressed in response to oxidative stress. The mitochondrial gene MT-RNR2 codes for Humanin, a 24 amino acid peptide. Humanin reduces oxidative stress in cardiac myocytes by decreasing the activity of Complex I of the Electron Transport Chain.\textsuperscript{110} Humanin also prevents apoptosis by forming a complex with Bax, preventing translocation from the cytosol to the mitochondrion.\textsuperscript{111} The long non-coding RNA MALAT1 protects against H\textsubscript{2}O\textsubscript{2}-induced oxidative injury by upregulating the transcription factor nuclear-factor-E2-related factor 2 (Nrf2), which coordinately regulates the expression of a number of anti-oxidant genes.\textsuperscript{112}
SPATA12 is anti-oxidant and, similar to MT-RNR2 and MALAT1, it also inhibits \( \text{H}_2\text{O}_2 \)-induced apoptosis. Another gene correlated with PSOx1 ATXN7 is an essential component of the STAGA (SPT3-TAF9-ADA-GCN5acetyltransferase) transcription coactivator complex. ATXN7 knockdown experiments reveal that protein encoded by this gene is required for complex stability and coactivator function.\(^{113}\) The complex is organized into separate modules that have distinct functions including a histone acetyltransferase (HAT) and histone deubiquitinase (DUB) activities. The HAT catalytic subunit of the yeast complex is GCN5; and ATXN7 interacts directly with GCN5.\(^{113}\) GCN5 and other components of the complex have been shown in model organisms to be essential for the organism’s survival after the initiation of oxidative stress by hydrogen peroxide. The molecular interaction network within \textit{S. cerevisiae} shows that Gcn5 and other complex components evince physical and genetic interactions with several genes involved in the oxidative stress response including superoxide dismutases, glutathione synthase 2 and gluaredoxin 5.\(^{114}\)

![Figure 5](image-url)  
Correlation between Process Signature Ox1 and RNA expression level, Correlation between Process Signature Ox2 and RNA expression level.  
Figure 5: The correlation between process signature exposure and RNA expression level. A. Process Signature Ox1 B. Process Signature Ox2.
E. Correlation with germline variants and methylation

If PSOx1 is the result of the incorporation of oxidized dNTPs, the question arises: Are these oxidized species present due to increased oxidative stress within the nucleotide pool, decreased protection against this stress or inadequate DNA repair mechanisms? Once again, we can use PSOx1’s similarities to Signature IRb for insight. Chronic exposure to ionizing radiation is correlated with the increased expression of proteins involved in the protection of cells against oxidative stress. Among these proteins are glutathione transferase omega-1, peroxiredoxin 6, glutathione-S-transferase P, and manganese superoxide dismutase 2. Perusal of the DNA methylation data correlated with PSOx1 reveals that there are eleven regions within GSTO1 that are statistically significantly (FDR<0.05) associated with this signature suggesting that decreased levels of this protective protein may increase the concentration of oxidized dNTPs.

To protect against DNA damage due to oxidative stress, premature cellular senescence is induced; to wit cells are irreversibly arrested in the G1 phase of the cell cycle and are no longer able to divide. Sublethal doses of H2O2 induce cellular senescence, which is associated with CHEK2 activation and an irreversible growth arrest in G1. Methylation of two upstream islands in the CHEK2 gene is correlated with PSOx1. Inactivation of the Chk2 protein has been shown to decrease p21 expression, and extend the proliferative lifespan of cells consistent with a failure to activate p53. Another way in which the cells associated with PSOx1 may be avoiding the tumor suppression that occurs secondary to cellular senescence is by dysfunction of the histone acetyltransferase complex that includes ING4 and HBO1.

F. Process Signature Ox2

Process Signature Ox2 is dominated by G>T (C>A), G>C (C>G) and G>A (C>T) mutations
Haradhvala et al. have suggested that G>T (C>A) and G>C (C>G) somatic mutations in breast cancer mainly take place during DNA replication.\textsuperscript{86} These mutations are likely the result of the exposure of peroxynitrite (ONOO\textsuperscript{-}) or nitrosoperoxycarbonate (ONOOCO\textsubscript{2}-), which subsequently undergoes homolysis to produce carbonate radical (CO\textsubscript{3}\textsuperscript{-}) and nitrogen dioxide (\textsuperscript{\cdot}NO\textsubscript{2}).\textsuperscript{116} The primary products of ONOO\textsuperscript{-} oxidation and nitration of guanine residues in DNA are 7,8-dihydro-8-oxoguanine (8-oxoG), 8-nitroguanine (8-NO\textsubscript{2}-G), 2-aminomidazolone (Iz), and 5-guanidino-4-nitroimidazole (NI).\textsuperscript{116} 8-oxoG adopts the syn conformation at the active site of DNA polymerases which permits Hogsteen base pairing with adenine during DNA synthesis.\textsuperscript{117} Iz induces G to C mutations, which are thought to occur due to its ability to mimic the hydrogen-bonding face of C and to form a stable base pair with G.\textsuperscript{118} NI is responsible for G>C, G>A and G>T mutations. While the bypass efficiency by the replicative polymerases is low, error prone polymerases such as the Y family may be involved in translesion synthesis (TLS) of this lesion.\textsuperscript{116} In addition to the specific mutations observed, another clue to the involvement of ONOO\textsuperscript{-} is the mutation motif. Rather than taking place at the nucleobase with the lowest redox potential, peroxynitrite initiated mutation is dependent on solvent accessibility of guanine bases and the extent of double-stranded character.\textsuperscript{119} Oxidation of the \textit{Kras} protooncogene by peroxynitrate reveals preferential oxidation at AGC and GG\textsuperscript{C}\textsuperscript{119}, two of the motifs with the greatest number of mutations in PS\textsubscript{O}x2 (Figure 6). Similar results are observed for the sup\textit{F} gene after exposure to 2 mM ONOO\textsuperscript{-}.\textsuperscript{120}
Figure 6: Ionization potential [1] and codon mutation frequencies in PSOx2. G in the center of the trinucleotides $\text{GGC}$ and $\text{AGC}$ (red arrows) is most frequently mutated consistent with nitrosoperoxycarbonate mediated oxidation [2]

G. Correlation of Process Signature Ox2 with RNA expression

Pathway analysis of the genes correlated with PSOx2 shows them to be involved in the ‘cell cycle’, ‘mitotic cell cycle phase transition’, ‘DNA replication’, and ‘microtubule cytoskeleton organization’ pathways.

Minimization of the number of genes resulted in a set of 50 that can distinguish the intrinsic subtypes, which is known as the Prediction Analysis of Microarrays or PAM50 assay\textsuperscript{121}. The basal subtype, as defined by PAM 50, is clearly enriched for tumors with the PSOx2 mutation pattern (exposure level is 28.4 (95% CI = 4.4) in 160 basal tumors and 6.0±0.4 in 750 non-basal tumors, p-value <0.01), suggesting an etiologic link with this subtype of breast cancer (Figure 7).
H. Correlation of PSOx2 with somatic mutations

Somatic mutations were aggregated within each gene, which is referred to as the “gene burden”. The 28 genes with a gene burden most highly associated with PSOx2 exposure are presented in Table 8. Immediately apparent from a perusal of the list of genes in the table is the association of TP53 mutations, and mutations in the DNA repair genes BRCA1/2 and FANCI with PSOx2. In addition to TP53, a number of other genes on the list are involved in the G1 checkpoint. G1 checkpoint is where the cell cycles were arrested and DNA repair would be performed. Among the members of the phosphoinositide kinase-like kinase family of proteins that includes ATM, ATR, and DNA-PK is SMG1. SMG1 regulates the G1/S checkpoint by both p53-dependent and independent pathways. Loss of function mutations may mimic that of TP53 as SMG1 phosphorylates p53 at Serine 15 in response to ionizing radiation, which leads to the stabilization and accumulation of p53. Elevated levels of CDK2 and loss of inhibitory phosphorylation of the encoded protein, which are observed in SMG1 depleted cells, is hypothesized to result in an
inappropriately active Cyclin E/CDK2 complex and consequent oncogenic proliferation. Another gene involved in the G₁/S cell cycle checkpoint is USP34, a substrate of ATM/ATR. Knockdown of this gene results in a deficient early G₁ checkpoint.

I. Interaction between genes and PSOx2 and the effect on survival

We evaluated the interaction effects between the gene expression levels and the signature exposure levels on survival status. Clinical covariates of age and stage were included to control for potential confounding bias. The gene with the most significant interaction is SLC25A29. Kaplan-Meier survival curves for the high PSOx2 and low PSOx2 group, stratified by SLC25A29 expression level, are shown in Figure 8. In high SLC25A29 expression level, high PSOx2 group is significantly associated with survival (HR=2.9; 95% CI: 1.0, 8.8; p=0.045). In the low SLC25A29 expression level group, survival curves of high PSOx2 and low PSOx2 groups are not significantly different (p=0.15). SLC25A29 is the main transporter of arginine into mitochondria. Arginine is a substrate for mitochondrial nitric oxide synthase, and it appears that mitochondrial derived NO constitutes a major proportion of whole-cell NO. Peroxynitrite is formed by the reaction of nitric oxide with superoxide and mitochondria are the major intracellular source of the superoxide anion. Peroxynitrite can permeate cell membranes, has a half-life of 5-20 milliseconds and can participate in reactions up to one-two cell diameters (~10μm) from its site of formation, which suggests that mitochondrial-derived peroxynitrite can reach nuclear DNA.
Figure 8: Kaplan-Meier survival curves for high Process Signature Ox2 versus low Process Signature Ox2. P-values were calculated using log-rank tests. A. Tumors with high SLC25A29 expression level. B. Tumors with low SLC25A29 expression level.
IV. Discussion

A. Process Signature Ox1

PSOx1 is one of the two oxidative signatures and is hypothesized to be the result of the incorporation of 7,8-dihydro-8’-oxo-dGTP (8-oxodGTP), 7,8-dihydro-8’-oxodATP (8-oxodATP) and 2-hydroxy-2’-deoxyadenosine-5’-triphosphate (2-OH-dATP) into nascent DNA. The ovarian steroids and their metabolites are not considered to be classic mutagens. However, it is not likely to be simply a matter of hormone exposure but exposure in the context of predisposing molecular defects upending hormone metabolic homeostasis. Cavalieri and Rogan posited that the disruption at any of a number of points on the finely regulated catechol estrogen pathway leads to the initiation of breast cancer. These investigators suggested that catechol estrogens result in the formation of depurinating estrogen-DNA adducts. We also hypothesize the involvement of the catechol estrogens but in the production of H$_2$O$_2$ and consequent oxidation of the nucleotide pool. Although we cannot rule out an excessive production of catechol estrogens, the methylation and germline SNVs correlated with PSOx1 suggest that there is defective protection against oxidative stress and/or that tumor suppression provided by cellular senescence is non-functional.

B. Process Signature Ox2

Integrating all of the data we have developed, we advance the following as an hypothesis to explain how the mutational process responsible for PSOx2 may lead to the development of the basal subtype of breast cancer. Based upon our data (Table 1), genes involved in the DNA damage response and the G$_1$/S checkpoint appear inordinately vulnerable to oxidative damage. This may be due to a specific sequence context acting as a trap for long-range hole migration to
solvent accessibility, to chromatin organization in the vicinity of the genes, or to other variables. Loss of the G1/S checkpoint appears crucial, as at least 3 of 29 genes listed in Table 1, TP53, SMG1, and USP 34, participate directly or indirectly in the regulation of in this checkpoint. This loss makes cells dependent on the S- and G2 checkpoints to provide a pause for DNA damage repair. However, these checkpoints appear to be compromised as well in PSOx2. They are activated by the CHK1 pathway. CHK1 expression is correlated with PSOx2 (Pearson correlation coefficient = 29%, FDR= 3.0E-19) and it has been shown to be highly expressed in a study of triple-negative breast cancer specimens, 75% of which were also basal.129 One of the activities of Chk1 is the phosphorylation of the atypical E2Fs, E2F7 and E2F8. The principle function of the atypical E2Fs is to repress the expression of a set of genes critical for cell cycle progression.130 Phosphorylation of E2F7 and E2F8 provides a docking station for 14-3-3 proteins, which when docked inhibit the repression function of these transcription factors. The consequences are the expression of E2F regulated cell cycle genes, abrogation of cell cycle arrest and the prevention of apoptosis. The correlation between PSOx2 and RNA expression levels were calculated. The genes correlated with the expression of E2F7 (Pearson correlation coefficient = 26%, FDR= 1.9E-15) and E2F8 (Pearson correlation coefficient = 20%, FDR= 3.5E-09). A large fraction of the genes on these lists are E2F regulated genes and are the genes that define the basal subtype. Specifically, 34 of the 50 genes of PAM 50 classifier are in the list; 9 of the 10 genes of the classifier that are both highly expressed and associated with the basal subtype are cell cycle genes: UBE2C, MYBL2, CCNB1, RRM2, ANLN, KIF2C, CENPF, MKI67, CDC20;121 all of them are associated with the expression of E2F7 and E2F8, which in turn are highly correlated with Process Signature Ox2.

Most oxidative DNA damage, i.e., base and/or sugar damage, abasic sites, is rectified by base
excision repair (BER), which occurs at equivalent rates during both G₁ and G₂.¹³¹ Loss of the G₁/S checkpoint will lead to the accumulation of cells with oxidatively-damaged DNA in the S phase. In addition to their many other functions, Brca1/2 proteins protect against oxidative DNA damage.¹³² BRCA1/2 deficient cells display increased basal and induced levels of 8-oxo-dG. In addition to oxidized nucleotides and proteins, oxidants can create single-stand breaks that may be converted into S-phase associated double strand breaks (DSB), which can result in clastogenic DNA lesions in the absence of homologous repair. Additional consequences of loss of function BRCA1/2 mutations will be the lack of transcription-coupled repair (TCR) of 8-oxo-dG on the transcribed strand; the same lesion of the non-transcribed strand is repaired by BER.¹³³ This deficiency has been shown by DNA sequencing to lead to G to T transversions as we observe in PSox2. This brings us to a final question: Are G to T transversion mutations oncogenic? The Myh and Ogg1 double knockout mouse, which cannot repair oxidative lesions at guanine provides evidence in the affirmative. These animal models develop lung and ovarian tumors, lymphomas and a small minority of mammary tumors.¹³⁴

CHAPTER 4 Deep Learning Model for Cancer Type Classification

I. Introduction

Cancer is essentially a genetic disease.¹³⁵,¹³⁶ Changes in DNA sequences, which is termed as ‘driver’ mutations, confer the proliferation advantage in cells.¹³⁶ Advancement in DNA sequencing technologies has led to the systematic processing and analysis of genomes from numerous tumors varying in both cancer type and subtype.²¹,¹³⁷,¹³⁸ The increasingly readily available genomic data has resulted in the discovery of relevant functional mutations that affect genes and pathways important in various cancers.¹³⁹-¹⁴¹ Together, these genetic landscapes allow for cancer type
classification.

There are patients with unknown primary cancer, where the site of origin cannot be established in the examination of metastatic cancer cells.\textsuperscript{142} Accurate pathogenetically distinct tumor type classification and accurate site of origin prediction can help identify therapy target precisely, thus minimizes toxicity and maximizes treatment efficacy.\textsuperscript{143} In most cases, cancer classification has relied on tumor morphology and histology and gene expression profiling.\textsuperscript{143-146} Classification based on morphological appearance has limitations because the tumor with similar histological appearance can respond differently to the same treatment.\textsuperscript{147} Developing systematic and unbiased approaches for recognizing tumor type and sub-type remains essential.

More recently, cancer classification methods have utilized mutational profiles. Somatic mutations have been used to distinguish cancer types and the results yield an overall accuracy of 49.4\%, they concluded that using somatic point mutations alone as individual variables was not sufficient to classify cancer types.\textsuperscript{148} However, these methods have focused only on single somatic point mutation despite the numerous studies providing evidence of other mutation types being associated with cancer.\textsuperscript{24} Insertion and deletion mutations are associated with increased susceptibility to cancer.\textsuperscript{149,150} Furthermore, germline mutations have also been shown to be associated with cancer.\textsuperscript{23,24} It has been challenging to integrate these mutations for cancer-type classification. Due to the limitation of analysis power, traditional statistical learning methods, such as Bayesian classifier,\textsuperscript{151} SVM\textsuperscript{152} and KNN\textsuperscript{153}, are not expert in handling the high-dimensional features. That raises the challenge to study all types of mutations together. They usually require onerous highly customized feature engineering before classification. Thus, they cannot generalize well and is expert labor intensive.

Conventional learning algorithms rely heavily on data representations, which are typically
designed by domain expertise or human engineers. The amount of required human efforts make it
difficult to scale the study due to the complexity in genome sequences, whereas deep learning can
be unsupervised once set in motion.\textsuperscript{25} Deep learning has recently emerged based on big data, the
power of parallel computing and sophisticated algorithms are exponentially more efficient than
the conventional models in learning intricate patterns from high-dimensional raw data with little
guidance.\textsuperscript{25,154-156} Typically, convolutional neural networks (CNNs) computes convolution on
small regions by sharing parameters between regions,\textsuperscript{157} which allows training models on large
DNA sequences.
Recent work has explored the application of CNNs in the raw DNA sequence without defining
priori features. DeepBind\textsuperscript{158} was proposed to predict specificities of DNA and RNA binding
proteins using raw sequence reads, and was reported to outperform the state-of-the-art methods.
DeepSEA\textsuperscript{159} is a CNNs-based tool to learn a regulatory sequence code from chromatin-profiling
data, and to predict noncoding variants’ functional effects. DanQ\textsuperscript{160} is a hybrid convolutional and
bi-directional long short-term memory recurrence neural network for non-coding function
prediction; the study reported that there is a 50% precision-recall relative improvement compared
the related models in the area. DeepCpG\textsuperscript{161} is a CNNs developed to predict methylation states from
low-coverage single-cell methylation data; the study reported that methylation states and sequence
motifs associated with changes in methylation levels were accurately identified.
Inspired by the successful applications of deep learning models in genomics data, in this study, we
proposed a model to utilize deep learning for disease classification using exome sequencings
(DeepCues). Specifically, a CNNs model to study DNA sequences for cancer type prediction is
investigated. In addition to tumor sequence, we also investigated whether germline DNA sequence
is informative for cancer classification. Furthermore, we are also interested in identifying a subset
of genes that are most relevant for each cancer type.

II. Methods

A. Data sources and data processing

Germline and somatic mutations from 4,174 samples across seven major cancer types were obtained from the Cancer Genome Atlas (TCGA)\(^{21}\). The following cancers were analyzed: brain (GBM), breast (BRCA), colon (COAD), kidney (KIRC), lung (LUAD), prostate (PRAD), and uterine (UCEC). To obtain germline variants, aligned sequencing data from blood or adjacent normal tissues were recalibrated, and variants were called using HaplotypeCaller in GATK package using assembly hg19. SnpEFF was used for functional annotation.\(^{88}\) Variants annotated with moderate effects are defined as missense mutations and inframe shifts and variants annotated as high effects are defined as nonsense mutations. Somatic mutations from matched samples were obtained directly from TCGA. In total, 4,600 virtual machines were utilized for 119,000 CPU hours to complete these tasks. We identified 45,119,052 germline mutations and 957,115 somatic mutations from the 4,174 matched samples (Table 8).

Table 8: The number of samples of each cancer and the corresponding number of germline variants and somatic mutations. Variants annotated with moderate effects are defined as missense mutations and inframe shifts and variants annotated as high effects are defined as nonsense mutations.

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Sample Size</th>
<th>Germline Moderate</th>
<th>Germline High</th>
<th>Somatic Moderate</th>
<th>Somatic High</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM Glioblastoma Multiforme</td>
<td>763</td>
<td>9572 (55)</td>
<td>550 (3)</td>
<td>94 (30)</td>
<td>23 (6)</td>
</tr>
<tr>
<td>BRCA Breast Invasive Carcinoma</td>
<td>959</td>
<td>10911 (71)</td>
<td>641 (5)</td>
<td>65 (10)</td>
<td>16 (3)</td>
</tr>
<tr>
<td>COAD Colon Adenocarcinoma</td>
<td>420</td>
<td>10331 (241)</td>
<td>621 (15)</td>
<td>293 (79)</td>
<td>81 (18)</td>
</tr>
<tr>
<td>KIRC Kidney Renal Clear Cell Carcinoma</td>
<td>332</td>
<td>10882 (124)</td>
<td>634 (8)</td>
<td>51 (6)</td>
<td>14 (1)</td>
</tr>
<tr>
<td>LUAD Lung Adenocarcinoma</td>
<td>730</td>
<td>9717 (37)</td>
<td>555 (2)</td>
<td>217 (15)</td>
<td>43 (3)</td>
</tr>
<tr>
<td>PRAD Prostate Adenocarcinoma</td>
<td>440</td>
<td>9744 (53)</td>
<td>558 (4)</td>
<td>34 (22)</td>
<td>7 (3)</td>
</tr>
<tr>
<td>UCEC Uterin Corpus Endometrial Carcinoma</td>
<td>530</td>
<td>10894 (103)</td>
<td>650 (7)</td>
<td>645 (152)</td>
<td>160 (28)</td>
</tr>
</tbody>
</table>
B. Feature generation

Recently, the gene coding direction has been described to be informative for mutational profiles.\textsuperscript{86} Sequences were annotated as positive if the gene is encoded on the reference strand and as negative if on the complementary strand. To account for directionality, gene sequences used for feature generation were obtained from the RefSeq database.\textsuperscript{90} Sequences with no mutations in our cohort were excluded from analysis. These sequences were then combined to form a consensus matrix (Figure 9A). This consensus matrix is consisting of 24,286 transcripts and was used as a base matrix. The average length of these sequences were 3,375 bases. For individuals, the germline variants identified in the matched normal were embedded into the consensus matrix, forming a germline matrix (Figure 9B). Once a germline matrix was formed for individuals, somatic mutations were then embedded in germline matrix, forming a germline and somatic matrix (Figure 9C).
Figure 9: Feature generation for proposed models (A) The transcript sequences were retrieved from RefSeq and were formed as a consensus matrix (B) One’s germline variants were embedded in the consensus matrix, forming a germline matrix for each sample. The brown dots are the germline variants that one has, including polymorphisms, deletions, and insertions. As an illustration, single nucleotide polymorphisms are identified and embedded in transcript A, E, and H, in-frame shift deletion is embedded in transcript B. In-frame shift insertion is embedded in transcript C. Frame shift deletion is embedded in D, and a frameshift insertion is embedded in transcript E. For this sample, the transcript F and G remained the same. (C) One’s somatic mutations were embedded in the germline matrix (from B), forming a germline and somatic matrix. The green dots are the somatic mutations that one has, including SNVs, insertions, and deletions. As an illustration, the tissue gained somatic mutations in transcript A, E; gained a stop loss in transcript F; and gained a deletion that shifted the frame in transcript G.

It has been suggested that mutations prefer certain codons and the distance between amino acid changes have been described. Moreover, the position within the codon where the mutation occurs will determine if the expressed mutation is nonsynonymous or potentially synonymous. To incorporate codon information into our model features, one hot encoding was applied with
every three nucleotides and was encoded as a binary unit. The combination of four nucleotides (A, C, T, and G) results in a vector with 64 dimensions to represent each codon combination.

**C. Model Structure**

A convolutional framework that composes of multiple layers was used in our study (Figure 10). The framework has three components: input layer (Figure 10A), encoder layer (Figure 10B) (multiple convolutional and dense layers), and fully connected layer (Figure 10C). The input layer in component (A) uses one hot encoding to represent each input sequence as a N*64 binary matrix, with N equals the number of codons and 64 is the number of codons. Therefore, the input can be considered as a 1-D sequence with 64 channels. Component (B) is an encoder layer to encode the input to a lower dimensional vector. The encoder component contains a sequence of convolutional layers with six output channels and a fully connected layer for each output channel. Therefore, a vector of six outputs is generated by the encoder for each input sequence. One convolutional layer is composed of one 1-D convolutional layer followed by a Leaky Rectified Linear Unit (LeakyReLU) as activation function and an average pooling layer. The number of convolution layers is determined by the transcript length N and the kernel size for average pooling layer. Kernel size six is used for the average pooling. We have N codons in each transcript, therefore, we will have \( \log_6 N \) convolution layers for each transcript. Component (C) is a fully connected layer with k outputs for k diseases. The inputs of component (C) are the combinations of products from the component (B) generated under the sequence of transcripts. With the average of 3,375 bases in the transcripts, the encoder layer would have an average of 3~4 convolution layers. To note, we set the following parameters for our model: the number of input channels for the encoder layer: 64; the convolution kernel size: 3, the output channel size of the encoder layer: 3; learning rate: 0.001;
batch size: 32; number of learning epochs: 30. We used cross entropy loss as the loss function and Adam algorithm as the optimizer.

D. Model evaluation and relevant gene discovery

A training set, validation set, and testing set were created by randomly splitting the samples using...
a 7:1:2 ratio, respectively. Parameters were trained using the training set and chosen using the validation set. Precision, recall, and F-measure were calculated for each cancer type using the testing set. To compare the performance of our models to other conventional methods for cancer classification, we applied penalized logistic regression and linear support vector machine (SVM)\textsuperscript{164,165}. The performance was also compared between the germline matrix alone and the germline/somatic matrix. Evaluations were repeated ten times with different initial seeds.

To reduce computational load, we selected for genes that have been implicated in cancer using a list of 719 consensus genes from the Catalogue of Somatic Mutations in Cancer (COSMIC). In our dataset, we found these consensus genes corresponded to 985 transcripts and used these transcripts to train and evaluate classifiers. The baseline model was also trained using germline variants and somatic mutations found in these selected transcripts. To discover potentially relevant genes not known to be implicated in cancer, we also applied a multinomial logistic regression model to the remaining transcripts using disease type as an output, and the number of mutations in each transcript as inputs to identify the 985 top ranked transcripts based on p-value. Classifiers were trained, and evaluation was measured using only known pathogenic transcripts and also using a combination of the known and unknown pathogenic transcripts. It has been demonstrated that features frequently ranked top in different training sets yields a robust set of predictive features with stability\textsuperscript{166} To obtain a gene list with reasonable stability, we repeated training the classifiers with random seeds and reported the top 20 most frequent transcripts in each replication.

III. Results

A. Classifier performance using known pathogenic transcripts

We first trained convolutional neural networks (CNNs) using the 985 known pathogenetic
transcripts and calculated overall accuracy for all seven cancer types. Using only the germline matrix as an input, we achieved an overall accuracy of 73.9% (SE=0.7%) (SE is standard error). Using the germline/somatic matrix as an input, we achieved an overall accuracy of 77.3% (SE=0.9%). To compare our method with other conventional cancer classification methods, we calculated baseline accuracies using logistic penalized linear regression and linear SVM. Logistic penalized linear regression resulted in an overall accuracy of 51.5% (SE=0.5%) and 65.7% (SE=0.4%) using the germline matrix and germline/somatic matrix, respectively. Linear SVM resulted in an overall accuracy of 49.4% (SE=0.5%) and 58.6% (SE=0.2%) using the germline matrix and germline/somatic matrix, respectively (Figure 11). We found that our method significantly (p<0.01) outperforms these methods.

Figure 11: Comparing Prediction accuracy between DeepCues and baseline models, including penalized logistic regression (LR) and support vector machine (SVM) with liner kernel.

We performed classification for seven types of cancer: GBM, BRCA, COAD, KIRC, LUAD, PRAD, and UCEC. For each type of cancer, we calculated precision, recall, and f-measure using either the germline matrix or the germline/somatic matrix (Table 9). Using only germline data, we found BRCA and COAD had the highest F-measure scores. Using both germline and somatic mutation data, we found BRCA, COAD, and GBM had the highest F-measure scores. Adding the
somatic mutation data, F-measures for BRCA, GBM, and UCEC increased significantly (p<0.05).

Table 9: Precision and recall for our proposed model. The experiment is replicated for 10 times and the number in parenthesis is standard error. The bolded number are those that significantly improved after adding somatic mutation information.

<table>
<thead>
<tr>
<th></th>
<th>Germline Matrix</th>
<th>Germline + Somatic Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>BRCA</td>
<td>81.9%(2.7%)</td>
<td>83.4%(4.2%)</td>
</tr>
<tr>
<td>COAD</td>
<td>85.9%(1.9%)</td>
<td>83.9%(2.1%)</td>
</tr>
<tr>
<td>GBM</td>
<td>73.0%(1.5%)</td>
<td>66.5%(4.3%)</td>
</tr>
<tr>
<td>UCEC</td>
<td>76.3%(4.9%)</td>
<td>62.7%(6.9%)</td>
</tr>
<tr>
<td>LUAD</td>
<td>64.8%(3.2%)</td>
<td>75.5%(4.7%)</td>
</tr>
<tr>
<td>KIRC</td>
<td>76.9%(2.5%)</td>
<td>71.5%(2.8%)</td>
</tr>
<tr>
<td>PRAD</td>
<td>70.8%(4.2%)</td>
<td>55.7%(6.0%)</td>
</tr>
</tbody>
</table>

B. Classifier performance using known and unknown pathogenic transcripts

Similarly, to the prior analysis, we applied the same model and integrated both the 985 known and the 985 additional non-pathogenic transcripts. Using only the germline matrix as an input, we achieved an overall accuracy of 82.68% (SE=0.6%). Using the germline/somatic matrix as an input, we achieved an overall accuracy of 80.0% (SE=0.9%).

We performed classification for the seven types of cancer. For each type of cancer, we calculated precision, recall, and f-measure using either the germline matrix or the germline/somatic matrix (Table 10). Using only germline data, we found BRCA and COAD had the highest F-measure scores. Using both germline and somatic mutation data, we found BRCA, COAD, and UCEC had the highest F-measure scores.
Table 10: Precision and recall for our proposed model. The experiment is replicated for 10 times and the number in parenthesis standard error. The bolded number are those that significantly improved after adding somatic mutation information.

<table>
<thead>
<tr>
<th></th>
<th>Germline Matrix</th>
<th>Germline + Somatic Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>BRCA</td>
<td>87.1%(2.8%)</td>
<td>91.9%(1.6%)</td>
</tr>
<tr>
<td>COAD</td>
<td>87.8%(1.7%)</td>
<td>94.0%(1.6%)</td>
</tr>
<tr>
<td>UCEC</td>
<td>90.0%(1.8%)</td>
<td>72.8%(6.8%)</td>
</tr>
<tr>
<td>GBM</td>
<td>88.3%(4.4%)</td>
<td>58.6%(5.2%)</td>
</tr>
<tr>
<td>LUAD</td>
<td>69.1%(4.9%)</td>
<td>79.9%(7.2%)</td>
</tr>
<tr>
<td>KIRC</td>
<td>81.9%(2.7%)</td>
<td>81.3%(2.0%)</td>
</tr>
<tr>
<td>PRAD</td>
<td>76.4%(6.1%)</td>
<td>67.3%(7.2%)</td>
</tr>
</tbody>
</table>

C. Relevant gene discovery

Using the coefficients derived from the fully connected layer, the model can be extended to prioritize genes that are relevant for each cancer type. The analysis was repeated 10 times with different initial seeds and the top 20 genes were selected in each replicate. The genes were then ranked by appearance frequency in all the replicates. We performed relevant gene discovery for breast cancer and reported the top 20 for each study. We identified genes using the germline matrix alone and the germline/somatic matrix using the known pathogenic transcripts (Table 11) and also the known/unknown pathogenic transcripts (Table 11). We also performed relevant gene discovery for the other cancers using only the germline matrix and the germline/somatic matrix.

Using Germline/Somatic genes, 8 of the top 20 genes overlap with the COSMIC top 20 genes for breast cancer. The high consensus rate (40%) validated that our method is effective in identifying the relevant genes. We have also identified relevant genes that are unknown for breast cancer (bold genes under the panel of 1970 transcripts; Table 11).
Table 11: The top 20 genes relevant genes with breast cancer derived from the 985 pathogenetic transcripts and the 1970 transcripts. The bold genes in the 985 transcripts are the ones found in COSMIC top 20 genes. The bold genes in the 1970 transcripts are the ones in the unknown transcripts

<table>
<thead>
<tr>
<th>985 transcripts</th>
<th>1970 transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Germline</strong></td>
<td><strong>Germline/Somatic</strong></td>
</tr>
<tr>
<td>TCF3</td>
<td>GATA3</td>
</tr>
<tr>
<td>FOXP1</td>
<td>APC</td>
</tr>
<tr>
<td>LEF1</td>
<td>RNF213</td>
</tr>
<tr>
<td>PAFAH1B2</td>
<td>PIK3CA</td>
</tr>
<tr>
<td>BLM</td>
<td>PPP2R1A</td>
</tr>
<tr>
<td>PPFIBP1</td>
<td>CDH1</td>
</tr>
<tr>
<td>PALB2</td>
<td>CHD2</td>
</tr>
<tr>
<td>MUC4</td>
<td>RUNX1</td>
</tr>
<tr>
<td>CCDC6</td>
<td>MSH6</td>
</tr>
<tr>
<td>LIFR</td>
<td>CNBD1</td>
</tr>
<tr>
<td>MITF</td>
<td>KMT2C</td>
</tr>
<tr>
<td>CCNE1</td>
<td>ARID1A</td>
</tr>
<tr>
<td>ZCCHC8</td>
<td>TBX3</td>
</tr>
<tr>
<td>TP53</td>
<td>KRAS</td>
</tr>
<tr>
<td>FOXO4</td>
<td>CHD4</td>
</tr>
<tr>
<td>EBF1</td>
<td>KMT2D</td>
</tr>
<tr>
<td>TERT</td>
<td>NRAS</td>
</tr>
<tr>
<td>GMPS</td>
<td>MAP3K1</td>
</tr>
<tr>
<td>CCNC</td>
<td>HLA-A</td>
</tr>
</tbody>
</table>

IV. Discussion

The development of high throughput sequencing technology has enabled the cataloging of large-scale genetic information. To help improve cancer diagnosis and targeted therapies, cancer type classification methods are continually being upgraded. Traditionally, the majority of classification methods based on DNA sequencing data has relied on studying single point somatic mutations
with various regression models. Mutations involving insertions and deletions as well as germline mutations have been widely ignored due to the high dimension problem. Given that many methods are already limited in their ability to study so many variables, it has been even more challenging to integrate these variables and study them interactively. To deal with these challenges, groups have proposed aggregating mutations on a gene level to be studied as a feature. Mutations within genes have also been proposed to be studied within a matrix as inputs for machine-learning methods. In our study, we have proposed a novel method, DeepCues, not limited to point somatic mutations by integrating all somatic mutations and germline variants, including INDELs, to be studied as inputs in an interactive manner. Convolutional Neural Networks (CNNs) were then applied to train classifiers for cancer type classification. Furthermore, we have included a fully connected layer for gene ranking by coefficients, thus it enables relevant gene discovery, and further can help characterize genes and pathways important for various cancers.

As a use case, we retrieved germline and somatic DNA sequencing data from matched samples across seven types of cancer and used DeepCues to perform cancer type classification. Using 985 known pathogenic transcripts as an input, we obtained 73.9% and 77.6% accuracy when using germline data alone and germline/somatic data, respectively. Following the integration of additional 985 unknown pathogenic transcripts into the model, we were able to increase overall accuracy to 82.7% and 80.0% for germline data and germline/somatic data, respectively. DeepCues was found to significantly outperform conventional methods used to perform cancer type classification \( (p < 0.01) \). Consistent with somatic mutations playing a large role in cancer, integration of somatic mutations together with germline data significantly improved overall accuracy \( (p < 0.01) \) using the 719 known consensus genes. Integration of somatic data significantly
increased accuracy for breast cancer (p<0.01), brain cancer (p<0.01), and uterine cancer (p=0.02), suggesting somatic mutations play a relatively larger role in these cancers. Conventional methods have been limited regarding germline variation and their interactive role in cancer due to a large number of variables and complexity issues. In our study, we were able to obtain reasonable accuracy performances using the germline matrix only as an input suggesting germline variation may be more important than previously reported based on prior methods.\textsuperscript{148} More specifically, we found that BRCA and COAD performed the best using only germline information, suggesting that these two cancers probably confers higher heritability compared to others. Studies have reported high familial heritability in breast and colorectal cancer too.\textsuperscript{171} Using a fully connected layer in our framework, we identified relevant both known and unknown pathogenic genes found using both the germline and germline/somatic data. For the 20 genes we have identified to be relevant for breast cancer, 40\% of the genes have been reported in the COSMIC top 20 genes for breast cancer.

Future development to better evaluate and assess our model will involve the inclusion of gene expression level, copy number variation, methylation, as well as including additional transcripts to be studied. Given that DeepCues is novel in its ability to utilize germline data in an informative manner, it will be of great interest and clinical impact to apply DeepCues to differentiate cancerous and non-cancerous samples. Disease classification not only allows for improved diagnosis and therapies but also allows research to understand a disease through identified groups of genes and related pathways. DeepCues uses generic sequencing data as inputs with little domain knowledge and feature preparation. With the abundance of genomic information available, we expect DeepCues can be used in a variety of disease settings to help profile diseases.
CHAPTER 5 Genomic Events in Benign Breast Biopsy

I. Introduction

While great strides have been made in the treatment of breast cancer, successful prevention remains elusive. Current breast cancer prevention strategies fall into one of three categories: lifestyle modification, surgical intervention, and chemoprevention\(^{172,173}\). These strategies have had, at best, limited success.

All tissues accumulate DNA mutations over time.\(^{176}\) Driver mutations have been identified in physiologically normal skin cells.\(^{176}\) Most of the mutations are repaired, many are inconsequential but a few may lead to or cause cancer\(^ {14}\). Central of this theory is that randomly distributed somatic mutations accumulate in normal cells before transformation.\(^ {177}\) The accumulated mutations are non-random and occur within sequence motifs. The motifs provide evidence from which we can infer the process that created the mutation.\(^ {3,5-7,19}\) Before there is any histologic evidence of cancer, histologically normal tissue and premalignant lesions have been demonstrated to contain molecular aberrations that are associated with malignancy. In normal tissues at high risk for breast cancer, such as normal breast tissue adjacent to breast cancer or the contralateral breast, are accompanied by aneuploidy, increased genomic instability, a wide range of gene expression differences, development of large cancerized fields, and increased proliferation.\(^ {26}\)

Therefore, with the goal of improving prevention strategies, we have designed a case-control study of benign breast biopsy tissues aiming at discovering early genomics events for cancer development. We have retrieved the tissue blocks from the benign breast biopsies of 135 patients who subsequently developed breast cancer (cases) and from 69 matched controls, who have not developed breast cancer (Figure 12). We hypothesize that the somatic mutations identified in the
cases are significantly different from those in the controls.

II. Methods

Benign biopsies have been reviewed and confirmed by the pathologist. Biopsy sections that were classified as “non-proliferative” (class 1) and “proliferative” (class 2) were retrieved from paraffin-embedded (FFPE) tissue blocks. The tissues have been micro-dissected, and total genomic DNA has been extracted from the laser capture microdissection (LCM) samples.

Figure 12. A. Design of the BBCAR study. B. Study Design illustrating the number of sequenced samples.

Whole exome sequencings were conducted with the sequencing depth of 80-100x, and 80-90 million sequencing reads were generated for each sample. Most biopsy samples lack matched
normal DNA, requiring us to develop a biopsy-only somatic mutation calling model. To develop and validate such a model, we collected 26 matched normal DNA for somatic mutation calling, and genotyped 23 benign biopsies using Infinium Exome-24 Kit. We evaluated multiple machine learning models and adopted multiple perceptron layer for somatic mutation prediction. The model predicted that 95% of the mutations called by our benign biopsy only pipeline are somatic mutations. To study the transitions from benign biopsy to cancer, we further sequenced ten cancer samples that matched to the benign biopsies. The details of methods are explained in below.

A. Study design

Using the NMEDW, we designed a case-control study to test the hypothesis that mutations present in the benign breast biopsies of women, who went on to develop breast cancer years later, are different from those who did not develop this disease, and these mutations could be used as markers of the risk of breast cancer development (Figure 12). The assumption underpinning the proposed research is that the breast tissue of the cases harbor genetic alterations predictive of the eventual development of breast cancer that were present years before the malignancy could be detected. Additionally, it is assumed that although a specific geographic area of the breast, i.e., the breast tissue in the archival block, is available for assessment, any genetic alterations identified that were predictive would predict generalizable, that is, bilateral risk. Lastly, it is assumed that the genetic alterations are not simply those that predisposed or led to the benign lesion. The second of these assumptions presumes a field cancerization that encompasses both breasts and that predates the detection of the malignancy.

Cases (n=135) are women who have undergone a breast biopsy with a benign result on histology that predates the diagnosis of breast cancer by at least one year. The median interval from benign
biopsy to the diagnosis of cancer is 7.3 (SD=4.4) years. Controls (n=69) are women who have not developed breast cancer matched for age (± 2 years), duration of follow-up, race, and histology (Table 12). Controls were verified for no cancer development at 08/14/2018. To rule out the germline variants, 16 and 10 matched saliva were collected for germline DNA extraction in the case and control group, respectively (Figure 12).

Table 12: Distributions of demographic data and tumor characteristics between the Case group and the Control group. Student’s t-tests were performed for continuous variables and Pearson’s Chi-squared tests were performed for categorical variables.

<table>
<thead>
<tr>
<th></th>
<th>Case (135)</th>
<th>Control (69)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>49.7 (9.9)</td>
<td>49.8 (9.6)</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Menopausal status N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>114</td>
<td>76 (56.3%)</td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>90</td>
<td>59 (43.7%)</td>
<td></td>
</tr>
<tr>
<td><strong>Class N (%)</strong></td>
<td></td>
<td></td>
<td>0.78</td>
</tr>
<tr>
<td>*Class 1</td>
<td>115</td>
<td>79 (58.5%)</td>
<td></td>
</tr>
<tr>
<td>*Class 2</td>
<td>75</td>
<td>51 (37.8%)</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>10</td>
<td>5 (3.7%)</td>
<td></td>
</tr>
<tr>
<td><strong>ER status N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>109 (80.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>23 (17.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>3 (2.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Follow-up years (SD)</strong></td>
<td>7.3 (4.4)</td>
<td>16.6 (5.4)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td><strong>Has matched germline (%)</strong></td>
<td>20 (14.8%)</td>
<td>6 (8.7%)</td>
<td></td>
</tr>
<tr>
<td><strong>Has matched cancer (%)</strong></td>
<td>10 (7.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*class 1/non-proliferative: “Non-proliferation” and “Benign, NOS”
*class 2/proliferative: “Ductal Proliferation without atypia”

**B. Sample collection**

Benign biopsies retrieved from the consented patients have been reviewed and confirmed by the collaborating pathologist. A trained interviewer contacted, consented, and completed a detailed questionnaire with the patient. The best-matched areas of non-proliferative and typical proliferative changes for the case-control pair were used for the analysis. Specimen retrieval, storage, and processing procedures were carefully designed with attention to blinding the
investigators. Ten 10-micron sections per sample were cut from formalin-fixed, paraffin-embedded (FFPE) tissue blocks. The matched areas of interest were isolated by laser capture microdissection (LCM). In detail, we took slides to Center for Advanced Microscopy (CAM) and used Zeiss palm microscope for micro-dissection and collection of areas of interest in a 500μl adhesive cap (AdhesiceCap 500 opaque -Zeiss order number 415190-9201-000).

C. Library construction and sequencing

Total genomic DNA has been extracted from the LCM samples. Tissue DNAs were extracted by Qiagen AllPrep DNA/RNA FFPE it (Cat. No. 80234). DNA concentration was measured by Nanodrop. Greater than 300ng of DNA has been isolated from 89% of the specimens processed to date. The concentration and quality of gDNA samples were first assessed using Agilent 4200 TapeStation. Then 100-200 nanograms of DNA per sample were used to prepare single-indexed cDNA library using SureSelectXT Human All Exon V6 (58Mb) (Agilent). The resulting libraries were assessed for its quantity and size distribution using Qubit and Agilent 2100 Bioanalyzer. Two hundred pico molar per liter pooled libraries were utilized per flow cell for clustering amplification on cBot using HiSeq 3000/4000 PE Cluster Kit and sequenced with 2×75bp paired-end configuration on HiSeq4000 (Illumina) using HiSeq 3000/4000 PE SBS Kit. A Phred quality score (Q score) was used to measure the quality of sequencing. More than 90% of the sequencing reads reached Q30 (99.9% base call accuracy). Average sequencing depth was 80-100x and there were 80-90 million sequencing reads per sample.

D. Parallel alignment of whole exome analysis

We adapted the most commonly used open source software for genome alignment and variant calling. Read alignment and variant calling were performed according to the Broad Institute’s
Genome Analysis Toolkit (GATK) best practices pipeline. Reads were aligned to the human reference genome (hg19) using Burrows-Wheeler alignment. Picard 2.6 was subsequently used to sort reads and mark duplicates. To reduce systematic errors, sorted BAM files were separately generated based on the sequence lane that the reads were generated. By doing so, various technical features that are associated with artifacts can be filtered during duplicate marking and base recalibration steps. Base recalibration was done using the GATK 3.6 using dbSNP build 138 as a training set. Mutations were called and filtered using MuTect2 in the GATK package. To compare different mutation calling tools, VarScan2 and VarDict were also employed for mutation calling. After mutations were called and exported in VCF (variant call format), we used the GATK tool VariantFiltration to filter mutations. Mutations were removed from the analysis using the following criteria: quality by depth <2; FisherStrand >60; RMSMappingQuality <40; MappingQualityRankSumTest <-12.5; ReadPosRankSumTest <-8. To capture recurrent technical artifacts, we generate a Panel of Normals (PON) for Mutect2 analysis using the sequenced 26 germline DNA. The PON is created by running the variant caller Mutect2 individually on the normal samples and combining the resulting variant calls with the criteria of excluding any sites that are not present in at least two normals. To further ensure a high precision call rate, we filtered mutations with read depth less than 20. After filtering, mutations were annotated using ANNOVAR and SNPEFF.

III. Results

A. SNP array genotyping

To evaluate the performance of called somatic mutations, 17 samples were separately LCM dissected, and the extracted DNA were genotyped using Infinium Exome-24 Kit, which covers
240,000 markers in a catalog of exome variants. To note, the quality control sample had a call rate of 99.34%. During the QC period, no samples failed restoration. For quality control, three samples were repeatedly genotyped twice, and R-square rates were calculated for the overlap. The reported R-square rates are 98.86%, 99.29%, and 99.39%. The high overlap rates indicate high stability of calling variants from our DNA that were extracted from FFPE. Genotyped variants from the 17 samples were mapped to genomic assemblies (hg19). To note, only the calls with GC score larger than 0.15 were retained.

Figure 13: Allele frequencies of the somatic mutations that were called in both array and whole exome sequencing

To illustrate, the somatic mutations called in both the SNP array and in the WES were presented in Figure 13. As we can see, most of the mutations identified have low allele frequencies. The coordinates that appear both in genotype array and somatic mutations called by MuTect2, VarScan2, or VarDict were retrieved. The allele frequencies derived from both technologies were compared. The overlap number between array and MuTec2, VarScan2, VarDict were 384, 963, and 124511 respectively. When the allele frequency difference was larger than a certain threshold,
we consider the somatic mutation call as a false positive. With different allele frequency cutoffs, we plotted the accuracy rate in Figure 14. In different thresholds, Mutect2 consistently has a better performance than VarScan2 and VarDict. When the cutoff is 25% (half of 50%), in other words, mutation calls with allele frequency difference larger than 25% between the two technologies were considered as wrong calls, the accuracy of using Mutect2 was 85.42%. Based on the consistent high accuracy rate, we decided to use MuTect2 for further studies.

Figure 14: Comparison of the called mutations between genotype array and WES. Somatic mutations were called using Mutect2, VarScan2, and VarDict. With different allele frequency difference cutoff, we obtained different accuracy. The accuracies were plotted.

**B. Predictive model for somatic mutations**

Only 26 (12.6%) samples contain matched germline DNA, requiring us to develop and test a predictive model for somatic mutation detection. In the previous section, we have learned that mutations identified by MuTect2 have higher accuracy than the other callers. In this study, MuTect2 was used to call somatic mutations from the 26 benign biopsies and matched normal germline DNA. To reduce the false positive call rates, mutations that appear in dbSNP with
ANNOVAR\textsuperscript{87} index files (after removing those SNPs $< 1\%$ minor allele frequency (or unknown), or mapping only once to reference assembly, or flagged in dbSNP as "clinically associated") and not in COSMIC database (version 80) were labeled as germline variants. The called mutations from the 26 matched samples were used as gold standard for the predictive model training and testing. The somatic mutations called from these 26 samples were randomly split to cross-validation set and holdout test set based on a 7:3 ratio.

Figure 15: Workflow to train and test a predictive model for somatic mutation identification
Table 13: List of features used in the predictive model for somatic mutation identification. ‘Allele frequency’ is the mutation allele frequency, ‘Ref depth’ is the read depth of the reference allele, ‘Fre cohort’ is the number of appearances of this mutation in the cohort. SNP common is a binary variable indicating that the mutation appears in the database of dbSNP after removing those flagged SNPs (SNPs < 1% minor allele frequency (MAF) (or unknown), mapping only once to reference assembly, flagged in dbSNP as "clinically associated"). ‘COSMIC’ is a binary variable indicating whether the mutation appears in the COSMIC version 80. The other features were derived from functional annotations by ANNOVAR.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele frequency</td>
<td>Polyphen2 HVAR pred</td>
<td>MetaLR score</td>
</tr>
<tr>
<td>Ref depth</td>
<td>LRT score</td>
<td>MetaLR pred</td>
</tr>
<tr>
<td>Fre cohort</td>
<td>LRT pred</td>
<td>VEST3 score</td>
</tr>
<tr>
<td>SNP common</td>
<td>MutationTaster score</td>
<td>CADD raw</td>
</tr>
<tr>
<td>ExAC</td>
<td>MutationTaster pred</td>
<td>CADD phred</td>
</tr>
<tr>
<td>COSMIC</td>
<td>MutationAssessor score</td>
<td>GERP score</td>
</tr>
<tr>
<td>SIFT score</td>
<td>MutationAssessor pred</td>
<td>phylodP20way mammalian</td>
</tr>
<tr>
<td>SIFT pred</td>
<td>FATHMM score</td>
<td>phylodP100way vertebrate</td>
</tr>
<tr>
<td>Polyphen2 HDIV score</td>
<td>FATHMM pred</td>
<td>SiPhy 29way logOdds</td>
</tr>
<tr>
<td>Polyphen2 HDIV pred</td>
<td>MetaSVM score</td>
<td></td>
</tr>
<tr>
<td>Polyphen2 HVAR score</td>
<td>MetaSVM pred</td>
<td></td>
</tr>
</tbody>
</table>

Tools predicting somatic mutations in matched normal free samples have been developed in other studies.\cite{183,184} However, the developed tools attempted to predict somatic mutations in tumor-only samples. Tools that have been developed and validated using the mutations derived from tumor samples cannot be applied to benign biopsies directly. The feature landscapes of mutations derived from tumors and benign biopsies could be different. For example, allele frequency derived in tumor samples are expected to be higher than allele frequencies in benign biopsies.\cite{176} In order to predict somatic mutations in the benign-only biopsies without matched normal DNA, in this study, we attempted to develop and evaluate a new predictive model to predict somatic mutations in benign biopsies. Various tools have been developed to predict the potential pathogenicity. These tools consider either the protein structure, population frequency, or evolutionary factors.\cite{185} Varies functional annotation and toxicity scores were derived from ANNOVAR,\cite{87} COSMIC (https://cancer.sanger.ac.uk/cosmic), dbSNP/common (https://www.ncbi.nlm.nih.gov), along with
intrinsic sequencing features, such as mutation allele frequency, depth of reference reads, number of appearance in the cohort. In total, 31 features (Table 13) were retrieved or developed to create the predictive model for somatic mutation identification. In practice, not all mutations were functionally annotated and resulted with missing scores. The missing data that appear in more than one feature could bring a challenge to some of the classifiers (e.g., logistic regression). In this study, considering that the features are a mix of continuous number, binary feature, and categorical variables, Multivariate Imputation by Chained Equations (MICE)\textsuperscript{186} is used to impute the missing values. In detail, 20 sets of data were imputed with the iteration equal to 20. Consolidating the 20 sets of data into one, mean number was calculated for the continuous variables, and mode was calculated for the categorical variables.

Using the derived features, we evaluated multiple linear and nonlinear machine learning models for mutation prediction. The germline variants were treated as predictive negative, while the somatic mutations were treated as predictive positive. Grid searches were applied to tune each model’s parameters within the five-fold cross-validation set. To reduce the risk of having false positives, precision is used as selection criteria for parameter tuning and model selection. Once the model was tuned, it was applied to the holdout test, and the precision, recall, F-measure, and AUC score were reported. The model with the highest precision was selected as our somatic mutation predictive model. In this study, the evaluated machine learning methods include penalized logistic regression (LR), linear SVM, random forest classifier (RFC), gradient boosted tree (GBT), k-nearest neighbor algorithm (K-NN), SVM with rbf kernel, and multiple layer perceptron (MLP).

During the evaluation (Figure 16), the MLP obtained the highest precision (95\%) (Table 14) in a holdout test and was selected as our predictive model for somatic mutation prediction. The tuned MLP model has two layers and each layer with 10 and 5 neurons respectively. Learning rate was
set as ‘invscaling’ and solver was set as ‘lbfgs’. The ‘logistic’ activation function was applied in the MLP model.

Figure 16: Performance of different machine learning models in the holdout test. Penalized logistic regression (LR); linear SVM; random forest classifier (RFC); gradient boosted tree (GBT); k-nearest neighbor algorithm (K-NN); SVM with rbf kernel; multiple layer perceptron (MLP).

Table 14: The number of mutations in the train set and test set. The germline variants were treated as predictive negative and somatic mutations were treated as predictive positive. Using tuned multiple layer perceptron model to predict mutations in the holdout test set, the AUC score, precision, recall, and F-Measure are reported.

<table>
<thead>
<tr>
<th></th>
<th>Train Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germline</td>
<td>10464</td>
<td>3315</td>
</tr>
<tr>
<td>Somatic</td>
<td>2817</td>
<td>1115</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td>98.7%</td>
<td></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>95.0%</td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>97.6%</td>
<td></td>
</tr>
<tr>
<td><strong>F1-score</strong></td>
<td>96.3%</td>
<td></td>
</tr>
</tbody>
</table>

C. Validate predicted somatic mutations

After tuning the MLP model, we applied it to predict germline variants/somatic mutations on the mutations derived from the 178 benign biopsies without matching germline DNA. In total, out of
the 93,653 mutations, 38,210 were predicted to be somatic mutations. To study and validate the predictive accuracy of somatic mutations we have obtained, we randomly selected and genotyped three samples for an evaluation study. Three samples were separately LCM dissected. The extracted DNA were genotyped using Infinium Exome-24 Kit. Genotyped probes with GCSCORE larger than 0.15 from the three samples were mapped to hg19 assemblies. Overlapped coordinates were retrieved, and the allele frequencies derived from both technologies were compared. In total, there were 250 overlap calls in the two technologies among the three samples. When the allele frequency difference was larger than a certain threshold, we considered the somatic mutation call as false positive. With different allele frequency cutoffs, we plotted the accuracy rate in Figure S5. When the cutoff is 25% (half of 50%), in other words, mutation calls with allele frequency difference larger than 25% between the two technologies were considered as wrong calls, the accuracy of our called somatic mutations was 80.0%.

![Somatic Mutation Validation](image)

Figure 17: Comparison of the called mutations between genotype array and WES in the cancer samples in the three validation samples. Somatic mutations were called using Mutect2 and our developed somatic mutation prediction model. With different allele frequency difference cutoff, we obtained the accuracy plot.
D. Mutation catalogues

Among the 204 samples, we have identified 36,801 somatic base substitutions and 2,283 small INDELs (Figure 18C). There is a substantial variation in the number between samples. Number-wise, there is not a significant difference between the case and control (Figure 18A). Among the top 10 mutated genes, the case group and control group shared common genes (MUC17, OBSCN, FLG2, GLTPD2, ABCA13). At a glance, there are some samples that contain larger proportions of nonsense mutations in the case group compared to the control group (Figure 18C).

Figure 18: Cohort and catalog of somatic mutations in 204 benign breast biopsies. A. Mutation classes in the case group and control group B. The top 10 mutated genes in the case group and control group. C. Catalog of base substitutions, insertions/deletions in the 204 benign breast biopsies.
E. Highly mutated genes in the case group

We sought to determine if there are any genes significantly differentially associated with the cases versus controls. We fit logistic regressions for each gene, using the case/control as output and using the mutated individuals as inputs.

Figure 19: Genetic aberrations that distinguish case and control. For each gene, the percentage of mutated individuals in the case and control were shown. The Onc is the known oncogenes; the TS are the tumor suppressor genes. The p-values were derived using the case/control as output and the mutated individual as inputs in logistic regression. The middle panel shows the synonymous versus nonsynonymous rate. In the right panel, each column is an individual, and the color represents the mutation class.

The p-values were derived from the fitted model and the genes with significant (p<0.05) p-values
were selected and sorted (Figure 19). Three genes cancer genes (FLG, GNAS, CTNNA2) appear in the list, and these genes are enriched with nonsynonymous mutations.

Figure 20: Position of mutational alterations in the protein structure of FLG, GNAS, and CTNNA2.
The mutations enriched in FLG and CTNNA2, mostly fall at the protein domains (Figure 20), which increases the mutations chance of being pathogenetic. Specifically, Catenin Alpha2 (CTNNA2) plays an important role in cell adhesion and differentiation, and it has been found with increased risk (OR=1.2) for breast cancer in a genome-wide association study.\textsuperscript{187}

**F. Recurrent mutations**

We also sought to determine if there are any variants significantly differentially associated with the cases versus controls (Figure 21). After annotation, there were 21,237 mutations identified that occur either in splicing site or exons. Only mutations that are enriched in the case group (larger percentage) were of interest, and 13,361 mutations were selected. Subsequent analyses were performed using these mutations.

![Diagram](image)

**Figure 21:** Schema for identifying mutations within the case group

A two-proportional z-test was performed, and a P-values were obtained. 8 mutations with P-value < 0.05 were retained. The five nonsynonymous mutations and three synonymous mutations were listed in the table below (Table 15).
Table 15: Recurrent mutations enriched in the case control

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Position</th>
<th>REF</th>
<th>ALT</th>
<th>Benign %</th>
<th>Control %</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC3A1</td>
<td>chr2</td>
<td>44531284</td>
<td>T</td>
<td>G</td>
<td>5.9%</td>
<td>0.0%</td>
<td>Nonsynonymous SNV</td>
</tr>
<tr>
<td>GSTP1</td>
<td>chr11</td>
<td>67353579</td>
<td>C</td>
<td>T</td>
<td>16.3%</td>
<td>2.9%</td>
<td>Nonsynonymous SNV</td>
</tr>
<tr>
<td>RGL3</td>
<td>chr19</td>
<td>11516007</td>
<td>T</td>
<td>G</td>
<td>19.3%</td>
<td>5.8%</td>
<td>Synonymous SNV</td>
</tr>
<tr>
<td>TRIOBP</td>
<td>chr22</td>
<td>38120429</td>
<td>T</td>
<td>C</td>
<td>7.4%</td>
<td>0.0%</td>
<td>Synonymous SNV</td>
</tr>
<tr>
<td>ZBTB16</td>
<td>chr11</td>
<td>114112911</td>
<td>T</td>
<td>C</td>
<td>6.7%</td>
<td>0.0%</td>
<td>Synonymous SNV</td>
</tr>
<tr>
<td>KCNQ2</td>
<td>chr20</td>
<td>62038378</td>
<td>A</td>
<td>T</td>
<td>11.9%</td>
<td>2.9%</td>
<td>Synonymous SNV</td>
</tr>
<tr>
<td>SEPN1</td>
<td>chr1</td>
<td>26136282</td>
<td>C</td>
<td>T</td>
<td>14.1%</td>
<td>4.3%</td>
<td>Nonsynonymous SNV</td>
</tr>
<tr>
<td>IRF7</td>
<td>chr11</td>
<td>615103</td>
<td>G</td>
<td>A</td>
<td>5.9%</td>
<td>0.0%</td>
<td>Synonymous SNV</td>
</tr>
</tbody>
</table>

G. Identify cancer genes in the matched cancer samples

The case group is defined as benign biopsies that developed breast cancer at least one year later after the biopsy. In the case group, we have retrieved ten matched cancer blocks. The same procedures as benign biopsies were performed, including LCM dissection, DNA extraction, library construction, sequencing, alignment, mutation calling, and variant filtering. Especially, due to the relatively higher allele frequency in tumor tissues, the read depth cutoff was adjusted to 10, instead of 20. This is a cutoff widely used in other studies. In total, 10402 mutations were identified in these ten cancer samples.

Of the 10402 mutations, 957 mutations were found both in the benign biopsies and cancer tissues. Inferring that these are the mutations found in the benign breast biopsies, and were carried along to the cancer tissues. The average allele frequencies for these mutations is 32.2% (SD=18.7%) in the benign biopsy and is 46.7% (SD=17.3%) in the cancer tissues. The increase of allele frequency suggests a clonal expansion. FAT1, CTNNA2, MYH11, ATR are among the top ten mutated genes are also are known cancer driver genes.
H. Mutational processes

Using the somatic mutations in the case and the control group, we have identified three mutational signatures in each group. In both case and control group, we have identified the “aging” signature (WTSI Signature 1b; Figure 22; cosine similarity score: 83.2% for the case and 83.0% for the control), which is the putative result of the hydrolysis 5-methylcytosine. We also identified the “mismatch repair” signature (WTSI Signature 1b; Figure 22; cosine similarity score: 80.5% for the case and 80.1% for the control). We have then identified a new signature for each group, compared to the one in the control group, the new signature in case group has more C>A with C at the 5’.

Figure 22: Mutational processes identified in the case group and the control group. Both groups have the aging signature and mismatch repair signature. We also identified a new signature in each group.

Using the case and control together, we studied the mutational processes. The methods used for
mutational processes derivation is an unsupervised process. Using the signatures derived, we tried to study the associations between the signature exposure and the phenotypes. In an association study, we found that the novel signature we have identified is significantly (p=0.007) associated with triple negative breast cancer (Figure 23). In the association study, we then controlled for the potential covariates of age, menopausal status, and biopsy class, the association remains significant (p=0.016). Suggesting that novel signature we have identified, is predictive of triple negative breast cancer.

Figure 23: The novel signature we have identified, is significantly associated with triple negative breast cancer

I. Copy number variation

Of the 26 samples that have matched normal DNA, we applied VarScan\textsuperscript{2180} to study somatic copy number variation using the benign breast tissue DNA and normal DNA. We then applied GISTIC\textsuperscript{2188} to study recurrent copy number variations. The cytobands we found in the case group, but not in the control group, were presented in Figure 24. The cytoband 19q13.33 has been reported brain cancer and prostate cancer\textsuperscript{189,190}. The identification of 19q13.33 suggested that the copy number duplication of this cytoband might be a predictive factor for breast cancer development.
Figure 24: Recurrent somatic copy number variation in the case group

IV. Discussion

Cancer is thought to occur as a consequence of the progressive accumulation over time of several somatic mutations. In this study, we have studied the genomic events in benign breast biopsy, which presage the development of breast cancer. The specimens used for this study were fixed in formalin, embedded in paraffin and have been in storage for years to over a decade. Formalin fixation and paraffin embedding degrade and chemically modifies DNA.\textsuperscript{191} In this study, we had rigorously evaluated the sequencing quality and mutation calling rates and concluded that we have obtained a set of valid mutations. We have developed a predictive tool to identify somatic mutations for the benign biopsy without matched normal DNA. Using the multiple layer perceptron as a classifier, we have achieved the precision of 95%. The later on validation using the array have also validated the range of this high precision.

Data from the Connecticut Tumor Registry for patients diagnosed with 1935-59 revealed the relative risk of the development of a second breast cancer was 2.9\textsuperscript{192}. These years predate the approval of tamoxifen in the United States (1977) and the first reports of trials of combination chemotherapy\textsuperscript{193}. The registry data showed risk to be a function of age, stage and length of time
since diagnosis of the initial breast cancer. The risk of a second breast cancer is higher in the ipsilateral breast, for example, in the cohort of patients in NSABP-B06 treated with lumpectomy alone, the recurrence rate was 35% at 12 years\textsuperscript{194}. The elevated risk, however, is not restricted to the ipsilateral breast; a review of the literature by Chen and colleagues suggest that the contralateral breast is at a 2 to 6-fold risk for the development of disease\textsuperscript{195}. Also supporting a filed effect in breast cancer are the observations that systemic therapies used to successfully treat breast cancer also reduce the incidence of contralateral breast cancer\textsuperscript{196,197}. Exhaustive 5mm bread loaf sections of mastectomy sections by Vaidya and colleagues showed that a majority of these mastectomies had multiple foci of disease\textsuperscript{198}. Foci were present outside the index quadrant in half of all specimens and in 79% of breasts with more than one focus of disease. In this study approximately half of the additional foci were within 2 cm of the primary; however, fully 10% were greater than 5 cm removed from the primary. Massively parallel targeted sequencing of ILC-LCIS pairs as well as multifocal LCIS has also been informative\textsuperscript{199}. While the majority of ILC-LCIS pairs are clonal, only 30% of the LCIS-LCIS pairs were. While the clonal LCIS-LCIS pairs could be located in the same or different quadrants, the non-clonal pairs were all located in different quadrants. Using the set of mutations, we have identified mutations that are significantly more frequently observed in cases, including some of the genomic events that are either in driver genes, or have been reported in cancer studies, suggesting that early genomic events take place in the benign biopsies. These are some findings of interest that would will be followed up.

The concept of field cancerization was proposed initially to account for the multiple lesions that develop in the aerodigestive tract exposed to the carcinogens present in cigarette smoke\textsuperscript{200}. In this case, the field arises from the exposure to the carcinogen. Conceptually, there are at least three possible fields that may underlie breast cancer. In addition to that from carcinogen exposure, a
field can also arise from a mutagenic or epigenetic phenomenon that enables the emergence of cells that have a growth advantage that enables the expansion of cell numbers. With an increased number of cells the probability of a second oncogenic event increases and when it occurs new subpopulations will emerge. The third field is that which occurs in the context of an hereditary gene mutation, that is, a field of genetically altered (predisposed) cells, that suffers additional genetic “hits”. The patterns of oncogenesis will be different depending on the field: multiple independent primaries are likely to arise in the field exposed to a carcinogen. A cell with a growth advantage as in the second type will establish a field where none existed a priori resulting in clonal outgrowth. Sequencing breast malignancies reveals that in the case of bilateral breast cancers the majority do not share common mutations when assessed by targeted NGS, which is consistent with independent lesions\textsuperscript{201} and a field created by carcinogen exposure. However 15\% of bilateral lesions did share mutations suggesting a clonal relationship. Targeted and whole genome sequencing of multifocal breast cancer has shown that the foci are clonally related and within individual foci there is evidence of clonal sweeps\textsuperscript{202}. These lesions arise in a field of cells that have a selective growth advantage. It should be pointed out that these fields are not mutually exclusive; there is no reason why more than one type of field could not be contributing as, for example, exposure to a carcinogen may provide certain cells with a growth advantage.

In this study, we have validated the quality of sequence data using SNP array. We developed a tool to differentiate the somatic mutations and germline variants. Using the identified somatic mutations, we finally identified mutations, genes, and copy number variation associated with case/control. Furthermore, we have identified a mutational signature in benign breast biopsy, which is potentially predictive of triple negative breast cancer.

**REFERENCE**
99


134. Xie, Y. et al. Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T...


<table>
<thead>
<tr>
<th>Citation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang, Q. et al.</td>
<td>Detecting somatic point mutations in cancer genome sequencing data: a comparison</td>
</tr>
</tbody>
</table>


