AD_____

Award Number: W81XWH-12-1-0080

TITLE: Advancing Our Understanding of the Etiologies and Mutational Landscapes of Basal-Like, Luminal A, and Luminal B Breast Cancers

INVESTIGATORS:

Arul Chinnaiyan, MD, PhD

CONTRACTING ORGANIZATIONS:

Fred Hutchinson Cancer Research Center
Seattle, WA 98109-1024

University of Michigan
Michigan Center for Translational Pathology
Ann Arbor, MI 48109-5940

REPORT DATE: Dec 2019

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command

Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
Dec 2019	Final	09/15/2012 - 09/14/2019

4. TITLE AND SUBTITLE
Advancing Our Understanding of the Etiologies and Mutational Landscapes of Basal-Like, Luminal A, and Luminal B Breast Cancers

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-1-0080

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Christopher Li, MD, PhD
E-Mail: cili@fhcrc.org

Arul Chinnaiyan, MD, PhD
E-Mail: arul@umich.edu

Peggy Porter, MD
E-Mail: pporter@fhcrc.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Fred Hutchinson Cancer Research Center (FHCRC)
PO Box 19024
Seattle, WA 98109-1024

University of Michigan
Michigan Center for Translational Pathology 5309 CCC
1400 E. Medical Center Drive
Ann Arbor, MI 48109-5940

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT: We conducted a five-year population-based case-case breast cancer study to identify how various breast cancer risk factors differ in their relationships to different molecular subtypes of breast cancer and to further characterize molecular differences between these subtypes To address the existing research gaps regarding the etiologies of different molecular subtypes of breast cancer we employed state of the art multidisciplinary approaches to advance our understanding of the epidemiology and mutational landscapes of basal-like, luminal A, and luminal B tumors. Our original goal was to recruit about 2,700 women in Western Washington who have been diagnosed with breast cancer to compare to 900 women who have never been diagnosed with breast cancer, but control ascertainment was unacceptably low, so on 8/25/15 we submitted a request to modify the SOW to drop the control group and replace it with an additional 80 to 100 ER+ cases. Participation in this research included a detailed telephone interview, collection of breast tissue and oral samples and medical record abstraction. Breast tissue samples were reviewed and tested at FHCRC and special tissue analyses were also performed at the Michigan Center for Translational Pathology. With rich epidemiologic and clinical data we are advancing knowledge regarding risk factors for the different subtypes of breast cancer. Another key component of this project is our deep molecular characterization of breast cancers that have recurred compared to those that have not recurred. This research may eventually be of help in developing clinically important insights and treatment protocols for future breast cancer patients.

15. SUBJECT TERMS
epidemiology, pathology, molecular subtypes of breast cancer, basal-like, luminal A, and luminal B tumors, breast cancer risk factors

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	50	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

I. INTRODUCTION

We conducted a five-year population-based case-case breast cancer study to identify how various breast cancer risk factors differ in their relationships to different molecular subtypes of breast cancer and to further characterize molecular differences between these subtypes. To address the existing research gaps regarding the etiologies of different molecular subtypes of breast cancer we employed state of the art multidisciplinary approaches to advance our understanding of the epidemiology and mutational landscapes of basal-like, luminal A, and luminal B tumors. Originally this study intended to include 900 newly diagnosed first primary triple negative (TN) invasive breast cancer cases,1800 randomly selected age-matched estrogen receptor positive (ER+) cases, and for comparison, a population-based control group of 900 women without breast cancer. The population-based control group was to be identified through Random Digit Dialing (RDD) using a system that automates the administration, execution, and tracking of the RDD process, but RDD control ascertainment did not go as planned due to significant changes in telephone equipment and practices, so the control response rate was unacceptably low. Consequently we made a request to modify the Statement of Work (SOW) on 8/25/15, which was approved. With the new SOW, we dropped the RDD control group and replaced it with an additional 80 to 100 estrogen receptor positive (ER+) cases. These additional cases provided us with more statistical power for our specific aims and increased the number of recurrences that we identified which is critical to Aim 5.

Study participants are residents of Western Washington and their participation in this study included a telephone interview that elicited detailed information on a variety of established and suspected breast cancer risk factors. At the end of the interview participants were asked to donate an oral tissue specimen for future genetic testing. Medical record reviews were also conducted to ascertain treatment and outcome (recurrence) information for cases. Additionally, tumor tissue specimens from all TN and ER+ cases were centrally reviewed at the Porter Lab and all tumors were tested for markers enabling us to identify the TN cases that are basal-like, and the ER+ cases that are luminal A vs. luminal B. Tumor tissue from confirmed basal-like, luminal A, and luminal B cases were sent to the Chinnaiyan Lab for mutational analysis. 133 cases from each group were used for discovery work, and the remainder were reserved for independent validation of promising candidates.

This research may eventually be of help in developing clinically important insights and treatment protocols for future breast cancer patients.

2. KEYWORDS:

Breast cancer, triple negative, ER positive, luminal A, luminal B, recurrence, risk factors, molecular profiling

3. ACCOMPLISHMENTS:

What were the major goals of the project?

OBJECTIVE

Our primary objectives are to characterize risk factors for and the mutational spectrums of basal-like, luminal A, and luminal B breast cancers by conducting the largest population-based study of basal-like and luminal B tumors to date. We are assessing how established and suspected breast cancer risk factors may be differentially associated with risk of these distinct tumor subtypes. Our preliminary data supported our hypothesis that substantial variations in these associations are present by molecular subtype. Given the known molecular heterogeneity of these tumors we also expect each subtype to have distinct mutational spectrums relevant to their etiologies. Determining risk factors and mutational signatures specific to basal-like, H2E, and luminal B cancers could impact clinical practice and public health in several respects: 1) Identifying modifiable risk factors for these cancers affords opportunities for prevention; 2) Development of targeted screening programs

for groups of women at high risk for these tumors affords an opportunity to diagnose these cancers earlier when they are more treatable; 3) Determining etiologic pathways relevant to these cancer subtypes at a population level can help inform the development of novel treatment and prevention strategies; and 4) Identifying subtype specific determinants of recurrence can help inform clinical decision making regarding treatment/follow-up for breast cancer patients.

SPECIFIC AIMS/GOALS

To address the existing research gaps regarding the etiologies of different molecular subtypes of breast cancer we will employ state of the art multidisciplinary approaches to advance our understanding of the epidemiology and mutational landscapes of basal-like, luminal A, and luminal B tumors. With a sample size that is substantially larger than any other study of basal-like breast cancer, we will address the following specific aims:

1. Identify and quantify risk factors for each of the most common molecular subtypes of breast cancer, basal-like, luminal A, and luminal B tumors, in a large-scale population-based study. We will evaluate all of the established breast cancer risk factors and additional hypothesized risk factors. Preliminary data indicate that oral contraceptive use and parity are more strongly related to risk of basal-like tumors.

2. Discover and validate the mutational landscapes of basal-like, luminal A, and luminal B tumors. Using next generation sequencing, we will characterize a large majority of point mutations, indels, amplifications/deletions and gene fusions from RNA and DNA isolated from formalin fixed paraffin embedded (FFPE) tissue specimens in a subset of our cases and use the remaining cases for validation.

3. Characterize the relationships between subtype specific risk factors and mutational signatures. The biological underpinnings of the stronger relationships between certain risk factors and basal-like breast cancer are unknown. Through this aim we will identify how various exposures influence the tumor genome.

4. Develop and validate risk prediction models unique to each breast cancer subtype incorporating epidemiologic and clinical data. Basal-like tumors are more likely to be interval detected rather than screening detected given the rapidity of their growth. The identification of women at high risk for these tumors, who may benefit from more frequent screening or imaging with complementary modalities, has the potential to identify these tumors earlier, when they are more treatable, and improve their prognosis.

5. Identify and quantify the relationships between various exposures and mutational changes on risk of breast cancer recurrence among patients with basal-like, luminal A, and luminal B tumors. While patients with basal-like and to a lesser extent luminal B tumors have poorer outcomes compared to luminal A patients, little is known about factors that influence prognosis.

What was accomplished under these aims/goals?

MAJOR ACTIVITIES

Task 1. Develop Interview Instrument, Other Study Materials, and Tracking System, Months 1-3, completed under Dr. Li's supervision at Fred Hutchinson Cancer Research Center (FHCRC):

a. Refinement of interview instrument, Months 1-2:
The telephone interview that will be used in this study has already been developed, refined through field testing, and is currently in use. Dr. Li and the study coordinator will review it again to see if any additional data elements should be collected based on any newly reported data relevant to our study aims.

Status: Complete

Development of other study materials, Months 1-2:
Several other materials will be prepared including: approach letters to potentially eligible cases and controls; an interview consent form; a HIPAA compliant authorization to access personal health information; tumor tissue (cases only), medical records, and pharmacy records release forms; and oral tissue specimen donation consent forms. Again, these forms will be adapted from our current work for use in this study.

Status: Complete, all the documents listed above were approved and used in the field.

Development of tracking system, Months 1-3:
We will modify the computerized tracking systems we currently use in our other studies to fit the specific needs of this project. This system will allow for up-to-date tracking of study progress and retrieval of information on any aspect of the study as needed.

Status: Complete, the tracking system is working efficiently.

Task 2. Obtain Institutional Review Board (IRB) Approval for Human Subjects Research, Months 1-6, completed under the supervision of Dr. Li at FHCRC and Dr. Chinnaiyan at the University of Michigan:

a. Obtain approval from the Fred Hutchinson Cancer Research Center and the University of Michigan IRB, Months 1-2:
Drs. Li and Chinnaiyan and their staff will lead the preparation and revision of documents necessary to obtain this approval.

Status: Complete, local IRB approval has been obtained at Fred Hutchinson and at the University of Michigan.

b. Obtain approval from the United States Army IRB, Months 1-4:
Drs. Li and Chinnaiyan and their staff will lead the preparation and revision of documents necessary to obtain this approval.

Status: Complete, DOD IRB approval has been obtained at Fred Hutchinson and at the University of Michigan.

Task 3. Training of Study Staff, Months 1-6, completed under Dr. Li's supervision at FHCRC:

a. Training of office based staff, Months 1-6:
Dr. Li will lead the finalization of all study protocols and the training of the study coordinator and program assistant on the implementation of these protocols.

Status: Complete, the study coordinator and program assistant have been trained. They worked closely with the Dr. Li to finalize and implement the study protocols and study staff has completed data collection.

b. Training of field staff, Months 4-5:
Under Dr. Li's supervision, the study coordinator will oversee the training of the study's field and random digit dialing (RDD) interviewers. To support this and on-going training, the study coordinator will develop a question by question (QxQ) specification manual that details the approach to be used for each question in the interview form.

Status: Complete, the QxQ and other training materials are finished. The telephone interviewers completed their training and finished their field work. The interviewers were experienced and well versed on our study specific procedures. They received extensive training on confidentiality, obtaining informed consent, administering the questionnaires, phone protocols, and collecting oral specimens. With the approved modified SOW, the RDD interviewer was trained to complete telephone interviews with the case group, since controls were no longer being recruited.

Task 4. Case Identification, Months 3-45, completed under Dr. Li's supervision at FHCRC:

a. Identification of potentially eligible cases from the cancer registry Months 3-45:
All potentially eligible cases will be ascertained through the Cancer Surveillance System (CSS), a population-based cancer registry covering 13 counties in western Washington State that participates in the NCI SEER Program.

Status: Complete, cases have been ascertained through the Cancer Surveillance System.

b. Verification of potentially eligible cases Months 3-45: The study coordinator will review the cancer registry abstracts and appended pathology reports to determine the eligibility of breast cancer cases. Lists of newly diagnosed cases will be generated twice each month from CSS files.

Status: Complete, CSS files for this study were downloaded and reviewed by the study coordinator each month.

c. Physician notification letters, Months 3-45:
The physicians of eligible patients will be notified in writing of our intention to contact their patients, to request updated contact information for these patients, and to solicit any reasons why patients may not be suitable for this study.

Status: Complete, per the Cancer Surveillance System (CSS), prior physician notification and physician permission is optional for research subjects diagnosed on or after January 11, 1992, so we only sent physician notification letters when we could not find selected tumor marker results or address and phone information for participants. However, living cases identified through the CSS are first approached with a letter from the CSS. As of October 2010, before any study can contact potential participants and their physicians, the CSS sends each potential participant a letter explaining that they are in the cancer registry, the registry's purpose, and that we are conducting a breast cancer study. The registry letter also gives potential participants a phone number to call within 10 days to opt out of the study. If the potential case does not call the registry in the specified time to opt out, a detailed approach letter and brochure explaining the study are sent. The CSS letter was used in the field for this study.

Task 5. Control Identification, Months 5-45, completed under Dr. Li's supervision at FHCRC:

a. Identification of controls, Months 5-45:
General population controls with no prior history of breast cancer will be identified through RDD using a system that automates the administration, execution, and tracking of the RDD process.

Status: Complete, RDD control ascertainment was halted in 2015 after a revised Statement of Work was reviewed and approved. The RDD interviewer was trained to complete telephone interviews with the case group.

Task 6. Approach to Study Subjects, Months 5-48, completed under Dr. Li's supervision at FHCRC:

a. Initial approach letter, Months 5-48:
Cases and controls will be approached about the study through a letter describing the study's purpose and procedures, and advising them that an interviewer will call soon.

Status: Complete, case enrollment has ended. An enrollment table is attached as Appendix 1. Per the approved modified SOW, RDD controls were no longer approached after 2015.

b. Initial telephone contact by a study interviewer, Months 5-48:
Within one week of the initial mailing, a trained interviewer will call the subject to answer any questions, verify eligibility, and schedule the interview. Then a letter confirming the appointment will be sent to subjects.

Status: Complete, interviewing for this study is done. An enrollment table is attached as Appendix 1. Per the approved modified SOW, RDD controls were no longer approached after 2015.

Task 7. Conduct of Interviews, Months 6-49, completed under Dr. Li's supervision at FHCRC:

a. Administration of study interview, Months 6-48:
All interviews will be conducted over the telephone. At the time of the interview the consent form will first be reviewed, any questions participants have will be answered, and consent will be obtained. We will enroll and interview a total of 900 TN cases, 1800 ER+ cases, and 900 controls. 700 TN cases and 900 ER+ cases are already been enrolled through another of Dr. Li's funded studies, so through this project we will enroll an additional 200 TN cases, 900 ER+ cases, and all 900 controls.

Status: Complete, the interviewers have completed all their assignments. They answered participant questions, obtained consents and conducted telephone interviews. Case enrollment is on target. A total of 1834 women with triple negative and estrogen receptor positive invasive breast cancer were enrolled in the study along with 129 RDD controls for a total of 1963.

b. Request for additional authorizations and release forms, Months 6-48:
HIPAA compliant authorizations to collect personal health information will be sought including:

i. A tumor tissue release so that specimens can be ascertained and centrally reviewed by the Porter Lab;

ii. A medical records release so we can review medical records including radiology reports from prior mammograms;

iii. A pharmacy records release that gives us permission to contact their usual pharmacies to verify and supplement reported medication use in future ancillary studies.

Status: Complete, the interviewers reviewed all these consents with the participants at the conclusion of the interview. The majority of the participants signed all the consents.

c. Request for oral tissue specimen, Months 6-48:
All subjects will be asked to donate an oral tissue sample using a provided Oragene kit. Participants will be asked to return their oral tissue sample along with all of their signed consent forms in a pre-paid envelope via the U.S. postal service.

Status: Complete, the interviewers requested an oral sample at the conclusion of the interview. The majority of the participants provided an oral sample.

d. Editing and coding of completed interviews, Months 6-48:
Interviewers will edit each interview within three days of their completion. Next, one of our staff members who has extensive editing experience will edit and code the completed questionnaires. Lastly, the study coordinator will conduct a final edit of all questionnaires and determine which subjects need to be re-contacted so that missing or incomplete data can be collected.

Status: Complete, the interviewers, editor, and study coordinator conducted edits and final reviews for all the completed interviews.

e. Validation of interview data, Months 7-49:
The study coordinator will review a random 5% sample of voice recordings from completed interviews as a quality controls check.

Status: Complete, 5% of completed interviews were reviewed for quality control and most interviewers received an excellent score.

Task 8. Tumor Tissue Review and Testing, Months 5-48, completed under Dr. Li's supervision at FHCRC:

a. Identification of tumor tissue specimens of interest, Months 5-48:
The study coordinator will review all cancer case abstracts assembled by CSS to identify the tumor tissue specimens of interest for this study including hospitals where they were ascertained, specimen characteristics, and specimen numbers. Tissue from surgeries post neo-adjuvant therapy will not be requested. This information will be entered into our study tracking system and used to generate hospital specific tissue request lists for this study that will be batched and requested from local hospitals quarterly by CERC staff under the study coordinator's supervision.

Status: The study coordinator finished identifying, prioritizing, and requesting tissue specimens. Tissues specimens for this study were batched by provider and requested quarterly.

b. Receipt and processing of specimens, Months 9-48:
We will request that all tissues be sent directly to the Porter Lab where Porter Lab staff will be responsible for tracking their receipt and organizing them for further processing by the Lab.

Status: We asked tissue providers to send the tissue specimens directly to the Porter Lab. Upon receipt, each specimen was checked-in, tracked and prepared for analysis by the Porter Lab staff.

c. Review and testing of specimens, Months 9-48:
Staff in the Porter Lab will conduct complete histopathologic reviews of each specimen obtained. They will also determine if the correct specimens have been received and if they are sufficient for further testing using immunohistochemistry (IHC). If additional tissue is needed, the Porter Lab informs the CERC group of this so any additionally surgical specimens potentially available can be requested. All ER+ cases will be evaluated for Ki-67 using IHC so that luminal A and luminal B cases can be distinguished from each other. All triple-negative cases will be evaluated for EGFR and cytokeratin 5/6 using IHC so basal-like cases can be identified.

Status: The review and testing of tissue specimens has been completed at the Porter Lab. The PIs and staff for the study and the Lab at FHCRC met monthly to review the study's progress and to address any issues or concerns.

Task 9. Review of Medical Records, Months 7-51, completed under Dr. Li's supervision at FHCRC:

a. Ascertainment and review of radiology reports, Months 7-51:
Medical records of breast cancer cases enrolled will be reviewed from date of diagnosis forward to ascertain information on breast cancer treatments and disease recurrences through the present date. Our medical record abstraction team lead by Ms. Zuanich has over 10 years of experience collecting data of this type from breast cancer patients in our region and interacting with each of the local hospitals and health care providers providing medical care to them.

Status: Complete, the medical record abstract team was trained and worked in the field and in the office to abstract data from electronic and paper medical records. Abstracting for this study is now complete. We completed medical record abstraction on 1646 participants.

Task 10. Integrative sequencing of specimens collected for this proposal, Months 7-60, completed under Dr. Chinnaiyan's supervision at the University of Michigan:

a. Process nucleic acids from FFPE specimens and ensure quality control, months 7-60: Extraction using QIAGEN FFPE RNAeasy and DNA protocols will be followed by analysis with the Agilent BioAnalyzer 2100 using RNA Nano and DNA 1200 reagents. If necessary DNA will be further fragmented using Covaris S2 adaptive focused acoustics shearing.

Status: Complete, we have optimized conditions to process FFPE specimens. All quality control measures are in place including determination of RNA quality, tumor content and genetic finger-print analysis to ensure sample integrity. In total, samples from 414 patients were sent.

b. Construct exome and captured transcriptome libraries. Sequence tumor and germline biospecimens (n=400 patients), months 7-60:
We will carry out whole exome sequencing of the tumor and matched germline specimens, and gene fusion assessment of the tumor transcriptome. Molecularly barcoded libraries for exome and transcriptome sequencing will be constructed using Illumina TruSeq protocols. Capture of the exome libraries and transcriptome libraries will be done using Roche EZ Exome v2 reagents and protocols. The three libraries for each patient will be multiplexed into a single lane and sequenced using Illumina TruSeq SBS v3 flowcells and reagents.

Status: We have extracted tumor DNA, normal DNA, and RNA from all 414 FFPE and/or blood samples that have been sent to Michigan. Library preparation using normal and tumor DNA and tumor RNA is complete for the basal-like (n=178) and luminal B cases (n=132). Library preparation for the 104 luminal A cases is complete. A total of 405 cases passed QC and were sequenced (tumor and normal OncoSeq panel and transcriptome).

c. Optimize our integrative sequencing approaches, incorporating improved methods and reagents for both increased speed and sequencing yield, months 7-60:
We will optimize and incorporate improved methods for library construction, such as transposon based addition of adapters, as they become available. We anticipate a continuing improvement in Illumina SBS reagents and procedures both increasing the speed and reducing the cost for each sample, as has been the case since the introduction of the technology.

Status: Complete, we have protocols in place that are optimized for carrying out the integrative sequence analysis. These include physically separate sample preparation module, reagent preparation module and post amplification module. Dedicated equipment is used throughout the sample and library preparation and downstream sequencing and analysis. We have incorporated the Illumina v4 chemistry and seen the anticipated improvement in both quality of sequenced bases and turnaround time on the sequencers. We have seen further improvements in depth and uniformity of coverage with the utilization of KAPA Hyper reagents for library construction.

d. Validation of mutations identified by integrative sequencing by targeted resequencing and QPCR methods, months 50-60:
Somatic point mutations and small indels pipelines have been validated by PCR amplification of the identified exons in tumor and matched normal DNA followed by PCR cleanup using Agencourt Ampure XP reagents and then sequencing on the ABI 3500 Genetic Analyzer using BigDye v3 protocols and reagents. Candidate gene fusions nominated from paired-end transcriptome sequencing and analysis will be validated by SYBR Green based QPCR using primers designed using Primer 3 to encompass the fusion junction on a panel of index case and control sample cDNAs.

Status: Completed for fusions and mutations.

Task 11. Bioinformatics analysis of integrative sequencing results, Months 7-60, completed under Dr. Chinnaiyan's supervision at the University of Michigan:

a. Analyze sequencing data to identify significant variants, months 7-60:
Using our in house developed tools, ChimeraScan, SNP detection and exome copy number, as well as available tools such as GATK and Snowshoes, we will analyze the patient sequence data.

Status: Primary analyses completed. All informatics analysis pipelines are in place. A web-based portal is in place that integrates data from a number of parallel analyses that are run by bioinformatics staff as well as coupled with a sample tracking LIMS system. Analysis pipeline consists of tumor content analysis, a SNP

"fingerprinting" QC analysis, somatic and germline mutation calls, insertions, deletions, copy number alterations, gene fusions, gene expression, and zygosity analysis from DNA and RNA sequence libraries. Results are integrated with lab and clinical data and public datasets such as COSMIC and Ensembl. Expression analysis of the basal cohort has been completed and a manuscript is submitted. Analysis of the luminal B and luminal A cohorts is nearing completion.

b. Interpret and translate sequence variants, months 7-60:
Identified mutations will be further analyzed for effects on reading frame and protein structure and function using analysis and prediction tools such as PolyPhen and PhastCons,

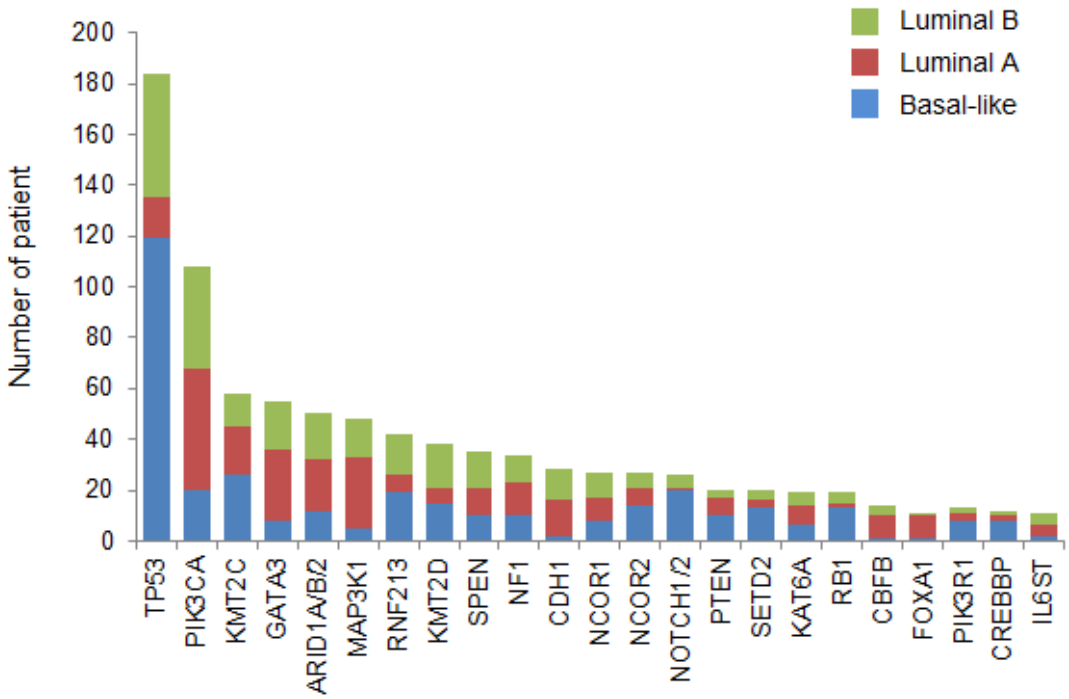
Status: Completed, Analysis tools are in place and are being used as samples undergo completed sequencing analysis. We have recently incorporated additional mutation evaluation tools, such as CADD (Combined Annotation Dependent Depletion) into the pipeline. We have fully implemented a bioinformatics pipeline for the presence of small indels in the sequence data and performed Sanger –based validation of the pipeline.

c. Optimize bioinformatics approaches, months 7-60:
We will continue to refine and test our bioinformatics pipelines and evaluate and incorporate new analysis tools and approaches as they are developed.

Status: Complete, we have a robust informatics pipeline in place supported by a web-based portal (see above). We have added LOH analysis for uncovering deleterious / tumor suppressor genes in the tumor samples. We have completed investigations on the feasibility of extracting reliable expression information from the FFPE derived transcriptome libraries. We have published an evaluation of the capture transciptome method we developed for use with FFPE and shown excellent performance in both expression data as well as gene fusion detection. (Cieslik, "The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing" Genome Research 2015.)

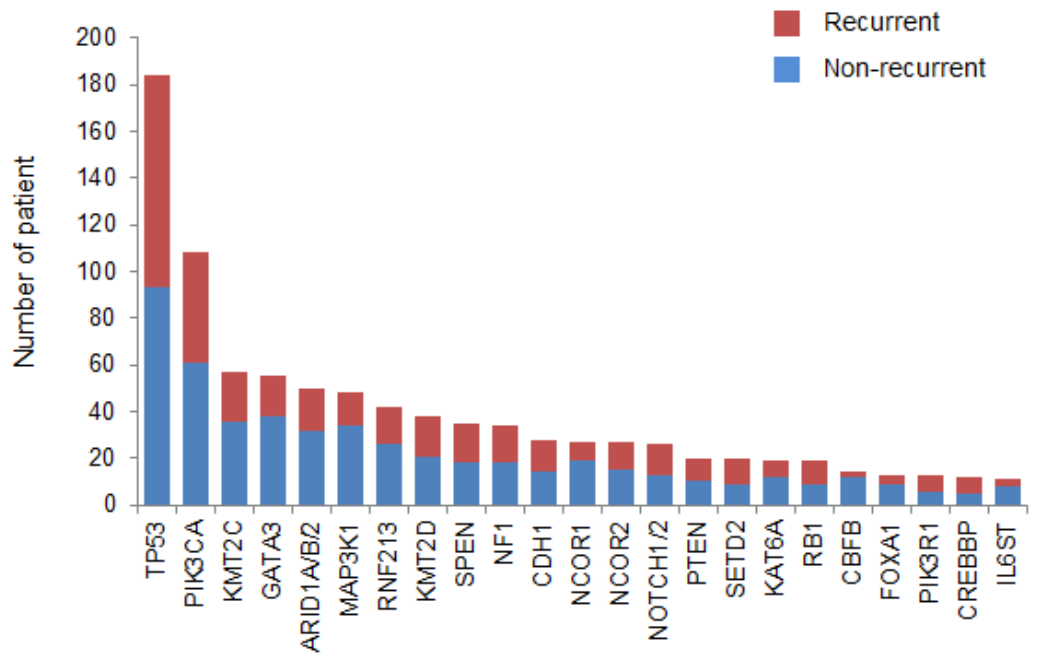
d. Incorporate primary sequencing results of the 400 patients in the discovery set with results from patients in the validation set, months 12-60:
Using a combination of Sanger based resequencing and pools of multiplexed PCR products resequenced on Illumina next-generation sequencing equipment, we will screen any recurrently mutated gene appearing in our 400 fully sequenced cases, across the entire set of available samples. Likewise, we will screen any functionally recurrent gene fusion across the complete sample set, using QPCR methods and further validation by Sanger sequencing of PCR products.

Status: <u>Complete,</u> complete sequence analysis for all accrued patients is completed. Primary analyses of exome and transcriptome data is now complete for all cohorts. Compilation of genetic events in all cohorts is complete. Mutation analyses for both germline and somatic variants are complete. Analysis of significant events delineating cancer subtype and recurrence is being finalized and manuscript preparation has begun. Further integration of mutation events with copy number events is nearly complete, e.g. combining mutation data of PTEN with copy number data of PTEN. Examples of mutation analyses and the gene level with regards to subtype occurrence and case recurrence data is shown as follows. (data for all 1700 genes is completed, not shown here.)



Distribution of somatic variants across subtypes

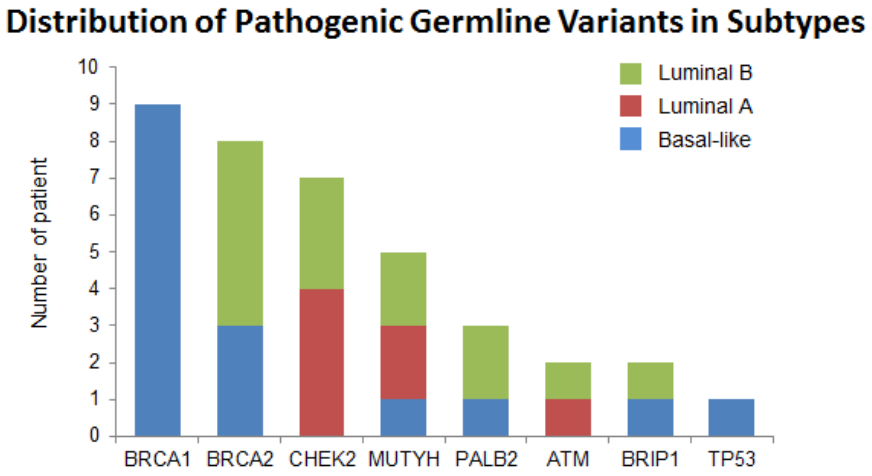
Significant differences are observed for TP53, GATA3, NOTCH, CBFB ($p < .05$). The distribution of all somatic variants in recurrent / nonrecurrent cases has been analyzed and results for the most frequently mutated genes are shown next.



Distribution of somatic variants by recurrence status

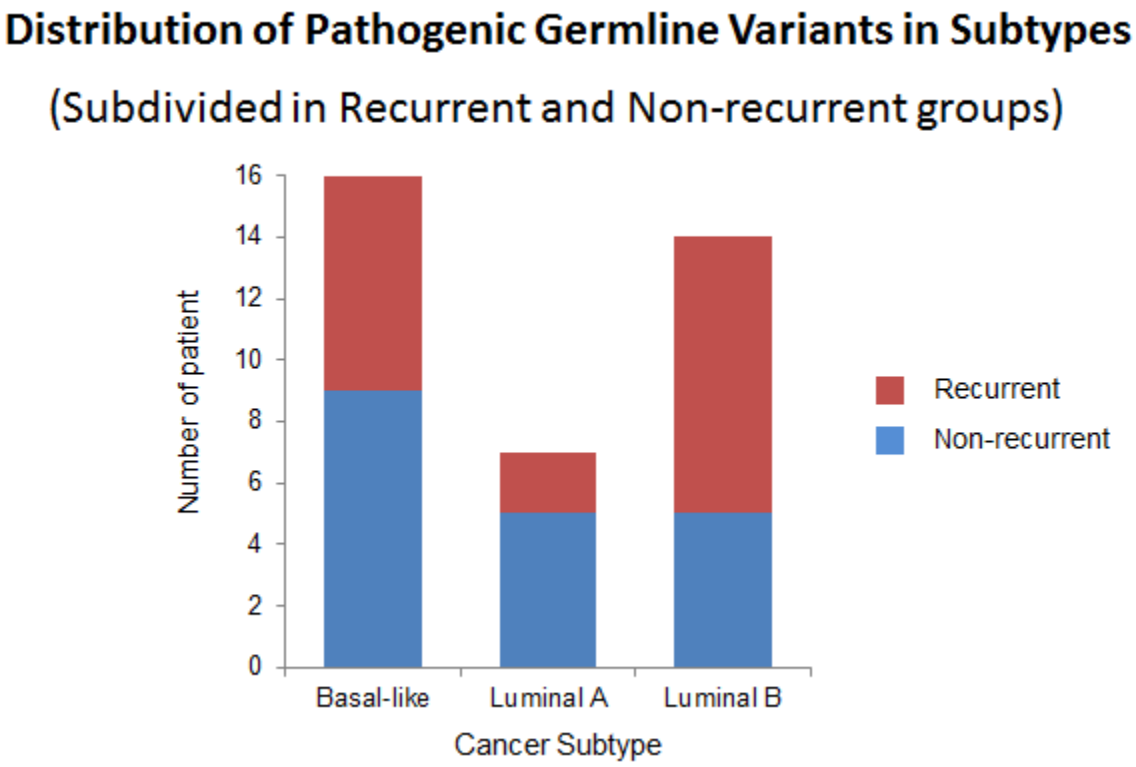
No single gene shows significant differences in the recurrent / non recurrent dataset. Analyses grouping genes into functional pathways for significance testing is ongoing, as well as combining copy number data for activating / inactivating events in pathways.

Analysis of germline variants for pathogenic alleles is complete. Sample results are shown next.



- 37 Pathogenic germline variants detected in 36 patients
- 28 out of 37 are biallelic events (75.6%), that include BRCA1 (9/9), BRCA2 (8/8), CHEK2 (4/7), MUTYH (1/5), PALB2 (1/3), ATM (2/2), BRIP1 (2/2), and TP53 (1/1)

Nearly 10% of the cases harbored pathogenic mutations in DNA repair and recombination pathways, with the large majority showing evidence of biallelic inactivation in the sample. Pathogenic germline variants were rarer in Luminal A subtype, and for no subtype was a significant difference observed for recurrent / nonrecurrent cases.



Task 12. Data Analysis and Manuscript Preparation, Months 52-60, completed under the supervision of Dr. Li at FHCRC and Dr. Chinnaiyan at the University of Michigan:

a. Data analysis, Months 52-60:
Dr. Tang, under the direction of Dr. Li, will lead the analyses of all epidemiologic data and the histopathologic data from the Porter Lab. Dr. Chinnaiyan and his team will lead the analyses of the mutational data generated in his lab. Data analysts from both groups will work together to realize the aims involving data collected from Dr. Li's field work and Dr. Chinnaiyan's lab.

Status: On-going, multiple analyses describing the results of the data generated from this study are currently in progress.

b. Manuscript preparation, Months 56-60:
Drs. Li and Chinnaiyan will lead the preparation of the multiple anticipated manuscripts that will describe the results of this study with the assistance of Dr. Porter and study staff.

Status: On-going. Multiple manuscripts have been published, the results of which are described below. Several others are either submitted or in development. In particular a manuscript describing the transcriptomic data from the basal-like cohort is currently in revision. Please see the Products Section below.

KEY RESEARCH ACCOMPLISHMENTS:

Summary of published manuscripts:

To date we have published four papers using data generated by this study. The results of these papers are summarized briefly below, and we refer you to the full publications for detailed descriptions of the methods, analyses, data tables, results, and conclusions.

1. **Chen L, Li CI, Tang MC, Porter P, Hill DA, Wiggins CL, Cook LS. Reproductive factors and risk of luminal, HER2-overexpressing, and triple-negative breast cancer among multiethnic women. *Cancer Epidemiol Biomarkers Prev*, 2016;25:1297-304.**

Summary: Reproductive factors are among the most well-established risk factors for breast cancer. However, their associations with different breast cancer subtypes defined by joint estrogen receptor (ER)/progesterone receptor (PR)/HER2 status remain unclear. We assessed relationships between reproductive factors and risks of luminal A (ER(+)/HER2(-)), luminal B (ER(+)/HER2(+)), triple-negative (TN; ER(-)/PR(-)/HER2(-)), and HER2-overexpressing (H2E; ER(-)/HER2(+)) breast cancers in a population-based case-case study consisting of 2,710 women ages 20-69 years diagnosed between 2004 and 2012. ORs and 95% confidence intervals (CI) were estimated with luminal A cases serving as the reference group using polytomous logistic regression. Earlier age at first full-term pregnancy and age at menopause were positively associated with odds of TN breast cancer (Ptrend: 0.003 and 0.024, respectively). Parity was associated with a 43% (95% CI, 1.08-1.89) elevated odds of H2E breast cancer, and women who had ≥3 full-term pregnancies had a 63% (95% CI, 1.16-2.29, Ptrend = 0.013) increased odds of this subtype compared with nulliparous women. Breast feeding for ≥36 months was associated with a 49% (OR 0.51; 95% CI, 0.27-0.99) lower odds of TN breast cancer. Our results suggest that reproductive factors contribute differently to risks of the major molecular subtypes of breast cancer. African American and Hispanic women have higher incidence rates of the more aggressive TN and H2E breast cancers and their younger average age at first pregnancy, higher parity, and less frequent breast feeding could in part contribute to this disparity.

2. **Chen L, Cook LS, Tang MC, Porter PL, Hill DA, Wiggins CL, Li CI. Body mass index and risk of luminal, HER2-overexpressing, and triple negative breast cancer. *Breast Cancer Res Treat* 2016;157:545-54.**

Summary: Triple negative (TN, tumors that do not express estrogen receptor (ER), progesterone receptor (PR), or human epidermal growth factor receptor 2 (HER2)) and HER2-overexpressing (H2E, ER-/HER2+) tumors are two particularly aggressive subtypes of breast cancer. There is a lack of knowledge regarding the etiologies of these cancers and in particular how anthropometric factors are related to risk. We conducted a population-based case-case study consisting of 2659 women aged 20-69 years diagnosed with invasive breast cancer from 2004 to 2012. Four case groups defined based on joint ER/PR/HER2 status were included: TN, H2E, luminal A (ER+/HER2-), and luminal B (ER+/HER2+). Polytomous logistic regression was used to estimate odds ratios (ORs) and associated 95 % confidence intervals (CIs) where luminal A patients served as the reference group. Obese premenopausal women [body mass index (BMI) ≥30 kg/m(2)] had an 82 % (95 % CI 1.32-2.51) increased risk of TN breast cancer compared to women whose BMI <25 kg/m(2), and those in the highest weight quartile (quartiles were categorized based on the distribution among luminal A patients) had a 79 % (95 % CI 1.23-2.64) increased risk of TN disease compared to those in the lowest quartile. Among postmenopausal women obesity was associated with reduced risks of both TN (OR = 0.74, 95 % CI 0.54-1.00) and H2E (OR = 0.47, 95 % CI 0.32-0.69) cancers. Our results suggest obesity has divergent impacts on risk of aggressive subtypes of breast cancer in premenopausal versus postmenopausal women, which may contribute to the higher incidence rates of TN cancers observed among younger African American and Hispanic women.

3. **Baglia ML, Cook LS, Tang M-T, Wiggins C, Hill D, Porter P, Li CI. Alcohol, smoking, and risk of HER2-overexpressing and triple-negative breast cancer relative to ER+ breast cancer. *Int J Cancer* 2018;143:1849-57.**

Summary: Epidemiological evidence is limited on how alcohol consumption and smoking are associated with risk of different subtypes of breast cancer, such as triple-negative (TN) and human epidermal growth factor receptor 2-overexpressing (H2E) breast cancers, which may have different etiologies from more common luminal (estrogen receptor [ER+]) breast cancers. In this population-based case-case study, we evaluated the association between alcohol, smoking, and risk of H2E and TN breast cancer, compared with ER+ breast cancers, among women aged 20-69 years. Using polytomous regression, associations between alcohol consumption, smoking, and breast cancer risk were evaluated in 909 ER+, 1,290 TN, and 489 H2E breast cancer patients, with ER+ breast cancer patients as the reference group. Current alcohol consumption at diagnosis was associated with a lower risk of H2E breast cancer (odds ratio = 0.74, 95% confidence interval: 0.58-0.92) relative to ER+ cancers. No difference in association was observed by menopausal status. No association between alcohol consumption and TN breast cancer relative to ER+ breast cancer was observed. Women who smoked did not have an altered risk of TN or H2E breast cancer, relative to ER+ cancer. Our results suggest that alcohol is associated with lower risk of H2E breast cancer relative to ER+ breast cancer. This study adds to the body of epidemiologic evidence that breast cancer etiology differs by breast cancer subtype.

4. **Chen H, Cook LS, Tang MC, Hill DA, Wiggins CL, Li CI. Relationship between diabetes and diabetes medications and risk of different molecular subtypes of breast cancer. *Cancer Epidemiol Biomarker Prev* 2019;28:1802-08.**

Summary: Type II diabetes and certain diabetes treatments have been observed to impact breast cancer risk. However, their associations with different breast cancer molecular subtype defined by estrogen receptor (ER)/progesterone receptor (PR)/HER2 status are unclear. We conducted a retrospective multi-center population-based case-case study consisting of 4,557 breast cancer cases to evaluate the impact of type II diabetes and diabetes medications on the risk of different breast cancer molecular subtypes [$ER^+/HER2^-$, $ER^+/HER2^+$, triple negative ($ER^-/PR^-/HER2^-$), and HER2 overexpressing (H2E, $ER^-/PR^-/HER2^+$)]. Using $ER^+/HER2^-$ cases as the reference group, we estimated ORs and corresponding 95% confidence intervals (CI) for each subtype using polytomous logistic regression. Compared with those without a diabetes history, women with type II diabetes had a 38% (95% CI, 1.01-1.89) increased odds of triple-negative breast cancer (TNBC).

Current and longer term recent metformin use (13-24 months of treatment within the 24-month period prior to breast cancer diagnosis) was associated with elevated odds of TNBC (OR = 1.54; 95% CI, 1.07-2.22 and OR = 1.80; 95% CI, 1.13-2.85, respectively). The odds of having a triple-negative rather than $ER^+/HER2^-$ breast cancer is greater for women with type II diabetes, and particularly for those who were users of metformin. This finding is supported by some preclinical data suggesting that diabetes may be more strongly associated with risk of triple-negative disease. Our study provides novel evidence regarding potential differential effects of type II diabetes and metformin use on risk of different molecular subtypes of breast cancer.

Summary of unpublished analyses/manuscripts in progress:

1. Relationship between obesity and risk of recurrence and mortality by breast cancer subtype

Patients with triple-negative and HER2-overexpressing breast cancer have increased risks of poor breast cancer outcomes, but little is known about the relationships between potentially modifiable lifestyle factors and risk of these adverse outcomes. In our study we have investigated the relationship between obesity and risk of both recurrence and mortality. Among only patients with ER-/HER2+ disease we have observed that obesity is associated with more than two-fold increases in risk of both recurrence and mortality as shown in the table below. In contrast, obesity did not increase risk of either of these outcomes among patients with other molecular subtypes of breast cancer. If confirmed, these results suggest that obesity is an important factor to consider in the treatment and clinical follow-up of patients with ER-/HER2+ breast cancer.

Table 1: Relationship between body mass index (BMI) and risk of recurrence and mortality by breast cancer subtype

Recurrence	**Luminal A**		**Luminal B**		**Triple negative**		**HER2-overexpressing**	
	n= 1,748 Time at risk= 8,313 years Number of events= 184		n= 268 Time at risk= 1,226 years Number of events= 32		n= 1,267 Time at risk= 4,046 years Number of events= 282		n= 487 Time at risk= 1,731 years Number of events= 67	
BMI (kg/m^2)	**Hazard ratio[a]**	**95% CI**	**Hazard ratio[a]**	**95% CI**	**Hazard ratio[b]**	**95% CI**	**Hazard ratio[c]**	**95% CI**
< 25	1.00	ref	1.00	ref	1.00	ref	1.00	ref
25-30	1.21	0.84-1.75	1.35	0.54-3.39	**0.62***	**0.45-0.85**	1.45	0.75-2.83
>30	1.08	0.76-1.55	1.25	0.51-3.05	0.93	0.70-1.23	**2.12***	**1.12-4.04**
Mortality	**Luminal A**		**Luminal B**		**Triple negative**		**HER2-overexpressing**	
	n= 1,809 Time at risk= 11,986 years Number of events= 127		n= 295 Time at risk= 1,899 years Number of events= 21		n= 1,344 Time at risk= 7,503 years Number of events= 241		n= 543 Time at risk= 3,158 years Number of events= 56	
BMI (kg/m^2)†	**Hazard ratio[a]**	**95% CI**	**Hazard ratio[a]**	**95% CI**	**Hazard ratio[b]**	**95% CI**	**Hazard ratio[c]**	**95% CI**
< 25	1.00	ref	1.00	ref	1.00	ref	1.00	ref
25-30	0.95	0.64-1.40	1.29	0.47-3.50	**0.66***	**0.50-0.88**	0.93	0.49-1.75
>30	1.20	0.85-1.69	1.18	0.47-3.01	0.89	0.67-1.17	**2.08***	**1.19-3.64**

* $p<0.05$

2. Antihypertensive use and risk of adverse breast cancer outcomes by molecular subtype

Hypertension is a common chronic condition impacting approximately one third of US adults. Diuretics, angiotensin-converting enzyme inhibitors (ACEIs), beta blockers (BBs), calcium channel blockers (CCBs), and angiotensin II antagonists (AIIAs) are medications commonly prescribed to treat hypertension. Findings on the association between common classes of antihypertensives and breast cancer outcomes have been mixed, with some studies finding an association with the use of certain antihypertensives and breast cancer outcomes, and others finding no association. We examined the relationships between hypertension and various antihypertensive medications and risk of three breast cancer outcomes, breast cancer recurrence, breast cancer-specific mortality, and all-cause mortality, by molecular subtype. As shown in Table 2, we observed that use of ACEIs was associated with increased risks of breast cancer specific mortality for both luminal and triple-negative patients, while calcium channel blocker use was associated with risk of breast cancer specific mortality among ER-/HER2+ patients. Our findings suggest that hypertension and some antihypertensive medications are associated with adverse breast cancer outcomes among women with certain molecular subtypes of breast

cancer. Future studies with adequate samples of women with less common subtypes of breast cancer are needed to confirm these subtype-specific associations and to provide more individualized clinical guidance for hypertensive women with breast cancer given the range of antihypertensive medications that are available.

Table 2. Hazard ratios and 95% confidence intervals for the risk of breast cancer recurrence, breast cancer-specific mortality (BCSM), and all-cause mortality (ACM) associated with hypertension and antihypertensive use, by molecular subtype

	Luminal			Triple-Negative			HER2-overexpressing	
	Recurrence[1]	BCSM	ACM	Recurrence[1]	BCSM	ACM	BCSM	ACM
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Hypertension								
No	ref	ref	ref	ref	ref	ref	ref	ref
Yes	1.4 (0.9, 2.2)	1.4 (1.0, 1.9)	1.7 (1.3, 2.1)*	1.2 (0.8, 1.6)	1.2 (0.9, 1.6)	1.4 (1.1, 1.8)*	0.9 (0.5, 1.5)	1.2 (0.8, 1.9)
Antihypertensive medications[2]								
Not current antihypertensive user	ref	ref	ref	ref	ref	ref	ref	ref
Current – any antihypertensive	1.4 (0.9, 2.2)	1.3 (0.9, 1.9)	1.6 (1.2, 2.1)*	1.2 (0.9, 1.7)	1.3 (1.0, 1.7)	1.5 (1.2, 1.9)*	1.4 (0.8, 2.4)	1.9 (1.2, 3.0)*
Current – diuretics	1.3 (0.8, 2.3)	1.1 (0.7, 1.7)	1.4 (1.1, 2.0)*	1.1 (0.7, 1.7)	1.1 (0.8, 1.7)	1.4 (1.0, 1.8)*	0.8 (0.4, 1.7)	1.6 (0.9, 2.8)
Current – CCB	†	1.5 (0.6, 3.6)	2.3 (1.4, 3.9)*	1.4 (0.8, 2.6)	1.3 (0.8, 2.3)	1.8 (1.2, 2.7)*	3.8 (1.4, 10.0)*	4.8 (2.1, 11.0)*
Current – BB	1.2 (0.6, 2.5)	1.2 (0.6, 2.2)	1.8 (1.3, 2.7)*	1.1 (0.7, 1.7)	1.2 (0.8, 1.8)	1.7 (1.2, 2.4)*	1.4 (0.7, 2.9)	1.6 (0.9, 3.1)
Current – ACEI	2.2 (1.2, 3.8)*	1.9 (1.2, 3.1)*	1.8 (1.2, 2.6)*	1.4 (0.9, 2.2)	1.6 (1.1, 2.5)*	1.7 (1.2, 2.4)*	1.4 (0.7, 3.0)	2.1 (1.2, 3.9)*
Current - AIIA	†	†	1.3 (0.7, 2.4)	1.8 (0.9, 3.8)	0.8 (0.4, 1.7)	0.8 (0.4, 1.5)	†	2.0 (0.7, 5.5)

1. Patients with Stage IV disease or a 2nd primary were excluded from recurrence models. Recurrence models restricted to patients with medical record review.
2. Current use of antihypertensive medications defined as use within the six months prior to one month before the date of diagnosis.

* Significant at $p<0.05$

3. Other epidemiologic analyses in progress

There are several other analyses that are currently in our queue for further development. These include evaluating a variety of other factors related to risk of recurrence/mortality by subtype including alcohol, smoking, cardiometabolic factors, cholesterol medications, bisphosphonate use, migraines, NSAIDs, and other commonly used prescription medications.

4. A prognostic signature for basal-like breast cancer integrating intrinsic and immunologic expression phenotypes

Basal-like breast cancer (BLBC) is a particularly aggressive intrinsic molecular subtype of breast cancer that lacks targeted therapies. There is also no clinically useful test to risk stratify BLBC patients. We hypothesized that a transcriptome-based signature characterizing BLBC tumors and their microenvironments may overcome these challenges. Through RNA sequencing of a matched set of 67 recurrent and non-recurrent BLBC tumors, we identified prognostic biomarkers through statistical-learning techniques. We developed a 21-gene signature we named BRAVO-DX that stratified patients into low-, medium-, and high-risk groups, with a 14% / 56% / 74% chance of recurrence, respectively. Biologically, the primary tumors of patients who developed a recurrence had increased growth-factor signaling and stem-like features, while non-recurrent tumors showed high lymphocyte infiltration with clonal expansion of T- and B-cells as well as anti-tumor polarization of macrophages. We validated BRAVO-DX in three independent cohorts where the signature was highly informative in identifying patients with disease recurrence (HR 6.79 [95% CI 1.89–24.37], HR 3.45 [95% CI 2.41–4.93], HR 1.69 [95% CI 1.17–2.46]). Together, these results indicate that phenotypic characteristics of

BLBCs and their microenvironment are associated with recurrence-free survival and demonstrate the utility of BRAVO-DX as an independent prognostic biomarker in BLBC. Pending further evaluation and validation, BRAVO-DX has the potential to inform clinical decision-making for BLBC patients as it identifies those at high-risk of rapidly progressing on standard chemotherapy as well as those who may benefit from alternative first line therapies.

Further details related to this analysis and its results can be found in the attached document which is a draft of our near final submission ready manuscript describing this work.

5. Mutation analysis of germline and somatic variants by subtype and recurrence status
See response to Task 11.d. above.

What opportunities for training and professional development has the project provided?
Nothing to Report

How were the results disseminated to communities of interest?
In 2020 we plan to send results letters to our study participants which will give a general overview of the study results and information about our recently accepted manuscript. This paper and future publications will be mailed to participants upon request.

What do you plan to do during the next reporting period to accomplish the goals?
Nothing to Report

4. IMPACT:
Through this project we have assembled a large cohort of highly characterized breast cancer patients with broad representation across the major molecular subtypes of breast cancer. With rich epidemiologic and clinical data ascertained through patient interviews and medical record reviews we are advancing knowledge regarding risk factors for the different subtypes of breast cancer. Indeed, the first publication from this study documents key differences in the relationship between diabetes and use of metformin and breast cancer risk by subtype. We observed that diabetes is associated with an increased risk of triple-negative breast cancer compared to women with ER+ disease. Further, use of metformin is also associated with a further increased risk of triple-negative breast cancer. Identifying potentially modifiable risk factors for these cancers and recurrences of these cancers affords opportunities for prevention. We are currently evaluating how factors such as obesity, alcohol consumption, and smoking are related to risk of recurrence within each of the major molecular subtypes of breast cancer through this study.

Another key component of this project is our deep molecular characterization of breast cancers that have recurred compared to those that have not recurred. Our goal is to identify patients up front at the time of diagnosis who have a high risk of recurrence and could benefit from alternative treatment strategies. Thus far we have primarily focused on data specific to triple-negative, basal-like breast cancer given that this is a particularly aggressive subtype. Based on our RNAseq data comparing 67 patients who recurred to 67 matched patients who are recurrence free we have developed a 21 gene signature that can accurately distinguish between patients with and without a recurrence. We have validated this signature in three independent, publicly available datasets of triple-negative patients including TCGA and METABRIC. A manuscript describing these results is currently in revision and if further validated it could have appreciable clinical impact. Specifically, it has the potential to identify triple-negative patients at the time of diagnosis who will rapidly fail standard chemotherapy. These patients should then be considered for more aggressive first line therapy such as being recommended for neoadjuvant chemotherapy and/or platinum-based chemotherapy. Alternatively, the patients with a very low risk of recurrence could potentially avoid some of the toxicities of the standard three-drug

chemotherapy regimen for triple-negative patients, and instead be candidates for a two-drug regimen that would avoid the cardiotoxicity of anthracyclines. Thus, further results from this study have the potential for identifying subtype specific determinants of recurrence that can help inform clinical decision-making regarding treatment/follow-up for breast cancer patients.

What was the impact on technology transfer?
Nothing to Report

What was the impact on society beyond science and technology?
Nothing to Report

5. CHANGES/PROBLEMS:

Changes in approach and reasons for change
Originally this study intended to include 900 newly diagnosed first primary triple negative (TN) invasive breast cancer cases,1800 randomly selected age-matched estrogen receptor positive (ER+) cases, and for comparison, a population-based control group of 900 women without breast cancer. The population-based control group was to be identified through Random Digit Dialing (RDD) using a system that automates the administration, execution, and tracking of the RDD process, but RDD control ascertainment did not go as planned due to significant changes in telephone equipment and practices, so the control response rate was unacceptably low. Consequently we made a request to modify the Statement of Work (SOW) on 8/25/15, which was approved. With the new SOW, we dropped the control group and replaced it with an additional 80 to 100 estrogen receptor positive (ER+) cases. The additional cases provided us with more statistical power for our specific aims, and in particular increased the number of recurrences that we identified which is critical to Aim 5. This modification did not impact the work or budget of Dr. Chinnaiyan's component of this project as there were never any plans to provide samples from controls to his lab for testing.

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

Changes that had a significant impact on expenditures.
Nothing to Report

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

Significant changes in use or care of human subjects
Nothing to Report

Significant changes in use or care of vertebrate animals
Nothing to Report

Significant changes in use of biohazards and/or select agents
Nothing to Report

6. PRODUCTS:

Publications, conference papers, and presentations
See below.

Journal publications
Chen L, Li CI, Tang MC, Porter P, Hill DA, Wiggins CL, Cook LS. Reproductive factors and risk of luminal, HER2-overexpressing, and triple-negative breast cancer among multiethnic women. *Cancer Epidemiol Biomarkers Prev*, 2016;25:1297-304.

Chen L, Cook LS, Tang MC, Porter PL, Hill DA, Wiggins CL, Li CI. Body mass index and risk of luminal, HER2-overexpressing, and triple negative breast cancer. *Breast Cancer Res Treat* 2016;157:545-54.

Baglia ML, Cook LS, Tang M-T, Wiggins C, Hill D, Porter P, Li CI. Alcohol, smoking, and risk of HER2-overexpressing and triple-negative breast cancer relative to ER+ breast cancer. *Int J Cancer* 2018;143:1849-57.

Chen H, Cook LS, Tang MC, Hill DA, Wiggins CL, Li CI. Relationship between diabetes and diabetes medications and risk of different molecular subtypes of breast cancer. *Cancer Epidemiol Biomarker Prev* 2019;28:1802-08.

Other publications, conference papers, and presentations
Nothing to Report

Website(s) or other Internet site(s)
Nothing to Report

Technologies or techniques
Nothing to Report

Other Products
Nothing to Report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Fred Hutchinson Cancer Research Center Employees						
Employee Name	Degrees	Project Role	Research Identifier	*Nearest person month worked	Contribution to Project	Funding Support
DeHart, Diane		Medical Record Abstractor	NA	2	Abstracted medical records	NA
Farr, Sheila		Interviewer	NA	4	Conducted participant telephone interviews	NA
Grogan, David		Programmer/ Database Manager	NA	2	Created and maintained study tracking systems and databases	NA
Hagen, Anna		Project Assistant	NA	2	Responsible for mailings and administrative duties	NA
Hicks, Joia		Medical Records Abstractor	NA	2	Abstracted medical records	NA

Li, Christopher	MD, PhD	Co-PD/PI	0000-0003-1543-0743	1	Responsible for oversight and implementation of all facets of the study conducted at FH	NA
Lin, Minggang	MD	Pathologist	NA	3	Reviewed pathology slides and interpreted IHC assays	NA
O'Brien, Heather		Medical Records Coordinator	NA	3	Medical record acquisition and oversight of abstractor work flow	NA
Porter, Peggy	MD	Co-Investigator	0000-0002-8271-401X	1	Responsible for oversight and implementation of tissue processing and testing	NA
Pride, Patty		Interviewer	NA	2	Conducted participant telephone interviews	NA
Ranney, Georgene		Interviewer	NA	3	Conducted participant telephone interviews	NA
Reitan, April		Program Assistant	NA	6	Responsible for tissue acquisition and administrative duties	NA
Tang, Mei	PhD	Statistical Research Associate	NA	4	Lead on the analyses of all epidemiologic data and histopathologic Porter Lab data	NA
Taylor, Sarah		Study Coordinator	NA	4	Coordination, QC, and supervision of data collection activities	NA
Thiel, Jodi		Project Assistant	NA	1	Provided general project assistance	NA
Tipton, Loni		Interviewer	NA	4	Conducted participant telephone interviews	NA
Wirtala, Kelly		Research Tech	NA	1	Responsible for tissue sectioning, IHC, and molecular assays	NA
Zuanich, Michelle		Medical Record Abstractor	NA	2	Abstracted medical records and provided abstract QC	NA

University of Michigan Employees

Employee Name	Degrees	Project Role	Research Identifier	*Nearest person month worked	Contribution to Project	Funding Support
Chinnaiyan, Arul	PhD	Co-PD/PI	0000-0001-9282-3415	1	Dr. Chinnaiyan will provide oversight for the processing of breast cancer FFPE sections for RNA/DNA, exome capture for whole exome and fusion analysis, as well as bioinformatics and mutation analysis. He will work closely with Dr. Li's team in associating mutations identified with clinical, pathological, and epidemiological factors.	NA
Chung, Fu-zon	PhD	Technician		3	Responsible for assisting in library generation and sequencing.	NA
Kalyana Sundaram, Shanker	MS	Bioinformatician	0000-0002-7309-3848	1	Mr. Kalyana-Sundaram has extensive expertise in bioinformatics and biology. He was one of the lead developers of www.oncomine.org. He will be responsible for analysis of the various multi-dimensional data sets generated as part of this proposal. He will be responsible for data assimilation and	NA

					annotation of clinical and pathology facts.	
Kunder, Komal Ramesh		Technician		2	Responsible for assisting in library generation and sequencing.	NA
Lonigro, Robert J	MS	Biostatistician	0000-0003-2124-172X	2	Mr. Lonigro is a staff biostatistician of the Cancer Center Biostatistics Core. He received training under Dr. Jeremy Taylor and Dr. Debashis Ghosh. Mr. Lonigro will be responsible for the biostatistical analyses on this proposal in the context of target nomination.	NA
Miller, Daniel A	BFA	Analyst		1	Mr. Miller will be responsible for developing and maintaining automated analysis pipelines to provide gene expression estimation and variant calling from transcriptome sequencing data.	NA
Mishler, Jeanmarie	MS	Technician		1	Responsible for sample extraction and library generation.	NA
Mohapatra, Pallavi	MS	Technician		2	Responsible for sample extraction and library generation.	NA
Ning, Yu	MM, MA	Technician		3	Responsible for assisting in library generation and sequencing.	NA
Robinson, Dan	PhD	Research Professor	0000-0002-2337-7439	1	Dr. Robinson will be responsible for and will oversee preparation of sequencing libraries, quality control, and provide expertise in genome biology. He is also first author on the paper in revision describing novel functional classes of gene fusions in breast cancer.	NA
Vats, Pankaj		Analyst		1	Pankaj Vats will be responsible for developing and maintaining automated analysis pipelines to provide gene expression estimation and variant calling from transcriptome sequencing data.	NA
Wang, Rui	MS	Technician		7	Ms. Wang has worked as research lab technician for over 17 years. She will extract nucleic acids from FFPE material and prepare cDNA from RNA samples. She also will prepare flowcells for sequencing using the cBot cluster generation equipment and assist in the running and maintenance of the HiSEQ 2000 equipment. She will also be responsible for monitoring DNA, RNA, and library quality using the Agilent 2100 BioAnalyzer.	NA
Wu, Yi-Mi	PhD	Research Scientist	0000-0002-3789-4445	2	Dr. Wu is a research scientist with over 10 years of laboratory experience. She completed her training at the University of California at Davis and she has expertise in numerous areas of molecular biology, including cancer genetics, retroviruses, cellular signaling pathways, and tyrosine kinase activation. Yi-Mi is an accomplished scientist and will provide important experimental expertise	NA

					in accomplishing the research proposal. Her skill and experience will guide the project's development, augment the project's research development, augment the project's research plan, and facilitate interpretation of experimental data.	

***Person months listed are average annual person months for entire 7-year project period.**

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Nothing to Report

What other organizations were involved as partners?
Organization Name: Michigan Center for Translational Pathology at the University of Michigan (UM)

Location of Organization: Ann Arbor, Michigan

Partner's contribution to the project:
The Michigan Center for Translational Pathology, under the direction of Dr. Arul Chinnaiyan, the study Investigator at UM, is responsible for AIM 2: Discover and validate the mutational landscapes of basal-like, luminal A, and luminal B tumors. The Center used next generation sequencing to characterize a large majority of point mutations, indels, amplifications/deletions and gene fusions from RNA and DNA isolated from formalin fixed paraffin embedded (FFPE) tissue specimens in a subset of our cases and used the remaining cases for validation. These contributions are detailed above in the *Accomplishment and Impact Sections*.

Financial support: Not Applicable

In-kind support: Not Applicable

Facilities: Michigan Center for Translational Pathology

Collaboration: See above

Personnel exchanges: Not Applicable

Other: Not Applicable

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS:
The sites at both Fred Hutchinson Cancer Research Center and the University of Michigan will submit this final report. Tasks are clearly marked with the responsible PI and research site in the *Accomplishments Section* above.

QUAD CHARTS:
Nothing to Report

9. APPENDICES:

Table 1 Study Enrollment

ADDITIONAL NOTES:
Nothing to Report

Appendix 1

Advancing our Understanding of the Etiologies & Mutational Landscapes of Basal-Like, Luminal A, & Luminal B Breast Cancers

Table 1: Final Enrollment

Status	Cases	Controls	**Total**
TOTAL ASCERTAINED	3109	253	**3362**
TOTAL INELIGIBLE	430	17	**447**
TOTAL ELIGIBLE	2679	236	**2915**
Review in Progress	0	NA	**0**
NON-PARTICIPANTS	845	107	**952**
Unable to Locate	146	1	**147**
Subject Refusal	648	106	**754**
Subject Refusal CSS Opt Out	51	NA	**51**
TOTAL IN PROCESS FOR INTERVIEW	0	0	**0**
TOTAL ENROLLED	1834	129	**1963**

A prognostic signature for basal-like breast cancer integrating intrinsic and immunologic expression phenotypes

Christopher I.Li[6,*,§], Yuping Zhang[1,§], Marcin Cieślik[1,2,3,9], Yi-Mi Wu[1,2], Erin Cobain[5], Mei-Tzu C.Tang[6], Xuhong Cao[1,2,4], Peggy Porter[7], Jamie Guenthoer[7], Dan R. Robinson[1,2], Arul M. Chinnaiyan[1,2,4,8,9,*]

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan 48109, USA
[2]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109, USA
[3]Depertment of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA
[4]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan 48109, USA
[5]Department of Internal Medicine Hematology/Oncology Division, University of Michigan, Ann Arbor, Michigan 48109, USA
[6]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[7]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
[8]Department of Urology, University of Michigan, Ann Arbor, Michigan 48109, USA
[9]Rogel Cancer Center, University of Michigan, Ann Arbor, Michigan 48109, USA
[§]These authors contributed equally to this work.
*co-corresponding authors

Key words: Basal breast cancer, triple-negative breast cancer, TNBC, BBC, prognostic biomarker, genomics, immune phenotype

Running Title: BRAVO-DX a prognostic biomarker for TNBC

Corresponding Authors:

Arul M. Chinnaiyan, M.D., Ph.D.
Investigator, Howard Hughes Medical Institute
American Cancer Society Professor
S. P. Hicks Endowed Professor of Pathology
Rogel Cancer Center
University of Michigan Medical School
1400 E. Medical Center Dr. 5316 CCGC
Ann Arbor, MI 48109-0602
arul@umich.edu

Christopher I.Li, M.D., Ph.D.
Member, Public Health Sciences Division
Fred Hutchinson Cancer Research Center
1100 Fairview Ave N, M4-C308
Seattle, WA 98109cili@fredhutch.org

Abstract

Basal-like breast cancer (BLBC) is a particularly aggressive intrinsic molecular subtype of breast cancer that lacks targeted therapies. There is also no clinically useful test to risk stratify BLBC patients. We hypothesized that a transcriptome-based signature characterizing BLBC tumors and their microenvironments may overcome these challenges. Through RNA sequencing of a matched set of 67 recurrent and non-recurrent BLBC tumors, we identified prognostic biomarkers through statistical-learning techniques. We developed a 21-gene signature we named BRAVO-DX that stratified patients into low-, medium-, and high-risk groups, with a 14% / 56% / 74% chance of recurrence, respectively. Biologically, the primary tumors of patients who developed a recurrence had increased growth-factor signaling and stem-like features, while non-recurrent tumors showed high lymphocyte infiltration with clonal expansion of T- and B-cells as well as anti-tumor polarization of macrophages. We validated BRAVO-DX in three independent cohorts where the signature was highly informative in identifying patients with disease recurrence (HR 6.79 [95% CI 1.89–24.37], HR 3.45 [95% CI 2.41–4.93], HR 1.69 [95% CI 1.17–2.46]). Together, these results indicate that phenotypic characteristics of BLBCs and their microenvironment are associated with recurrence-free survival and demonstrate the utility of BRAVO-DX as an independent prognostic biomarker in BLBC. Pending further evaluation and validation, BRAVO-DX has the potential to inform clinical decision-making for BLBC patients as it identifies those at high-risk of rapidly progressing on standard chemotherapy as well as those who may benefit from alternative first line therapies.

Introduction

Breast cancer is a heterogeneous disease with intrinsic subtypes revealed by histopathological and molecular profiling (Curtis et al., 2012; Network, 2012; Perou et al., 2000; Sørlie et al., 2001). Triple-negative breast cancers (TNBC), which lack expression of the estrogen receptor (ER) and progesterone receptor (PR) and do not overexpress HER2, represent 16% of all breast cancers and have a poorer survival rate (70% 5-year survival rate) compared to hormone receptor positive tumors (90-95% 5-year survival rates) (Dent et al., 2007; Sørlie et al., 2003, 2001; Trivers et al., 2009). Basal-like breast cancer (BLBC) is the most common type of TNBC, accounting for 70% of this subtype. Compared to other subtypes, BLBC is more common among younger, African American, and Hispanic women and is molecularly characterized by increased frequencies of germline aberrations in the homology-directed DNA repair pathway (Gonzalez-Angulo et al., 2011). Compared to hormone receptor-positive tumors, BLBCs are also associated with substantial genetic heterogeneity in terms of driver aberrations (Kawazu et al., 2017), passenger mutations and/or genetic instability, as well as clonal heterogeneity (Shah et al., 2012). Management of BLBC patients is associated with a number of major clinical challenges. BLBC patients have poor prognosis and recurrences that tend to occur within the first five years after initial

diagnosis, which is in contrast to hormone receptor positive disease where recurrences 5+years post-diagnosis are more common (Dent et al., 2007; Lin et al., 2012). At present, the standard of care for non-metastatic BLBC patients is primary local treatment (total mastectomy or lumpectomy with radiation) and either adjuvant or neoadjuvant anthracycline-taxane based chemotherapy. However, there is no BLBC specific test that differentiates patients who will rapidly fail standard treatment vs. those who will remain event free for five years or longer.

The host-tumor immune response has been explored as a prognostic indicator for TNBC and BLBC. More robust tumor immune cell infiltration and higher expression of immune-related genes have been associated with better clinical outcomes. For example, multiple studies have demonstrated that tumor infiltrating lymphocytes (TILs) are prognostic in early stage TNBC, with higher degrees of immune cell infiltration correlating with improved clinical outcomes (Adams et al., 2014; Denkert et al., 2014, 2010; Dieci et al., 2014; Liu et al., 2012; Li et al., 2016; Loi et al., 2013; Ono et al., 2012). More specifically, increased TILs are associated with increased likelihood of achieving pathologic complete response (pCR), an independent favorable prognostic indicator, following administration of neoadjuvant chemotherapy (Denkert et al., 2010; Li et al., 2016; Ono et al., 2012). Furthermore, for patients who do not achieve pCR following neoadjuvant chemotherapy administration, increased TILs in residual disease specimens have also been correlated with improved clinical outcomes (Dieci et al., 2014; Pelekanou et al., 2017b). TILs are also positively correlated with PD-1/PD-L1 expression in tumor and surrounding stroma (Brockhoff et al., 2018; Kitano et al., 2017; Pelekanou et al., 2017a), indicating that PD-1/PD-L1 inhibitors may prove effective in TNBC that is already primed with a robust tumor immune response. Indeed, the addition of pembrolizumab, a PD-1 inhibitor, to neoadjuvant chemotherapy significantly improved rates of pCR in patients with advanced triple negative disease (Nanda et al., 2017). Most recently, atezolizumab, a PD-L1 inhibitor has been approved for metastatic TNBC and is in the phase III IMpassion130 (NCT02425891) trial (Schmid et al., 2018), patients with PD-L1 have gained the largest benefits. Lastly, higher gene expression levels of T-cell receptor signaling pathway components, Th1 related cytokines, and B cell markers have also been correlated with increased likelihood of pCR after neoadjuvant chemotherapy and overall survival (Lee et al., 2015).

Use of immune checkpoint inhibitors, however, may not be appropriate in the upfront treatment of all patients with TNBC for three primary reasons: 1) not all patients will benefit from immunotherapy treatment, 2) some patients will have an excellent prognosis with chemotherapy alone, thus not justifying the potential added toxicity and costs associated with immunotherapy, and 3) chemotherapy-induced immunogenicity may augment the responses observed to anti-PD-L1 therapy. Given this, there is an urgent clinical need to risk-stratify patients with BLBC in ways that will meaningfully inform therapeutic decision-making. The purpose of this study was to use transcriptomic profiling to discover and validate

novel tumor-based biomarkers that may be useful for discerning BLBC patients who will from those who will not develop a cancer recurrence.

To address this clinical need, we performed RNA sequencing on tumor samples from 67 BLBC patients that developed a recurrence and 67 recurrence-free controls. These patients were selected from a large population-based prospective cohort of 1,408 triple-negative breast cancer patients 20-69 years of age diagnosed from 2004-2012. The prognostic signature developed through this discovery set was then validated in three independent cohorts of triple-negative patients including TCGA (Network, 2012), METABRIC (Curtis et al., 2012), and a cohort published by Gyorffy et al. (Györffy et al., 2010).

Results

Characteristics of discovery and validation cohorts

The discovery cohort was selected from a large population-based prospective cohort of the major molecular subtypes of breast cancer conducted in the greater Seattle-Puget Sound metropolitan area. The methods used in this study have been previously published (Baglia et al., 2018). Briefly, cases consisted of women 20-69 years of age first diagnosed with invasive breast cancer between June 1, 2004 and June 30, 2015 and identified through the population-based Surveillance, Epidemiology and End Results (SEER) program that served the 13 counties of western Washington state. Potentially eligible patients for this analysis (n=949) were identified from the 1408 patients with stage I-III TNBC (ER-/PR-/HER2-) and 949 were enrolled into this study (**Fig. 1A**). Of these 949 women, 218 developed a recurrence more than 6 months after their initial diagnosis and were identified as likely basal-like through positive staining for epidermal growth factor receptor (EGFR) and/or cytokeratin 5/6. This study was designed and powered to include 67 recurrent and 67 recurrence-free BLBC patients; out of the 76 patients with recurrence selected for inclusion, 67 had RNA of sufficient quality extracted and successfully underwent RNA-seq analysis. These 67 recurrent patients were matched to 67 recurrence-free patients 1:1 on age, diagnosis year, stage, and treatment (surgery, radiation, and chemotherapy). Patients were characterized as having a recurrence based on data extracted from medical records; both local and distant recurrences were included while deaths were excluded. Data on additional patient demographic characteristics, clinical factors, and established breast cancer risk factors were also collected through both medical record reviews and patient interviews (**Supp. Table 1**). Patients who received neoadjuvant chemotherapy were excluded because of the impact neoadjuvant therapy has on tumor gene expression from tissue collected post-chemotherapy. Given that recurrent and non-recurrent cases were matched on age, diagnosis year, stage, and treatment, the distributions of these variables were similar between the two groups. Of note, 97% of patients in both groups received adjuvant chemotherapy and 57%

received a total mastectomy. Publicly available expression data from three independent cohorts of TNBC patients were used for validation.

Transcriptomic Profiling of Formalin-Fixed Paraffin-Embedded Tissues in the Discovery Set

Pre-treatment formalin-fixed paraffin-embedded (FFPE) blocks of breast cancer tissue collected at the time of primary cancer directed surgery (lumpectomy or total mastectomy) or diagnosis (core biopsy) were attained from the hospitals where patients were treated. RNA was successfully extracted from the 134 patients included in this study and used to characterize the molecular and phenotypic characteristics of recurrent and non-recurrent basal breast tumors. Transcriptome profiling of FFPE specimens can be hindered by technical difficulties arising from RNA degradation (Sigurgeirsson et al., 2014) and cross-linking (Karmakar et al., 2015). Previously, we have established a hybrid cDNA exome-capture method followed by high-throughput sequencing, as an RNA sequencing protocol with improved robustness to RNA degradation and sample fixation (Cieslik et al., 2015). Therefore, we chose to perform "capture RNA-seq" on all samples in this study (**Supp. Table 1**). Still, the integrity of RNA extracted from specimens subject to fixation and long-term storage is expected to be highly variable. In order to identify samples of particularly low quality, we followed a data-driven strategy. First, we computed a battery of RNA-seq quality-control (QC) measures, including variables correlating with input RNA quality, library complexity, and *in vitro* generated molecular artifacts (Methods). Next, we clustered the RNA-seq libraries based on all of these computed QC measures (**Extended Data Fig. 1A**). This unsupervised analysis revealed that while no single QC parameter is sufficient to reliably indicate RNA-seq quality, the joint interrogation of multiple QC parameters identifies a set of samples with poor or marginal reads for many of them. We excluded these samples (n=14) from all following analyses. Next, to verify the accuracy and concordance of our transcriptomic platform, we tested whether the retained samples recapitulated intrinsic BC gene expression profiles (GEPs). We compared our samples to the TCGA breast cancer cohort comprising basal-like, luminal, HER2-amplified, and normal-like tumors profiled using poly(A)+ RNA-seq (**Fig. 1B**). We found that basal-like cancers from our cohort clustered together with the basal-like TCGA cancers, and 90.8% of the tumors classified as the PAM50 basal-like intrinsic subtype (Paquet and Hallett, 2015; Perou et al., 2010) (Methods). Next, we verified that recurrence status was not confounded by technical covariates. Critical parameters determined by specimen selection criteria (e.g. tumor purity [**Fig. 1C**] [(Aran et al., 2015) or stromal admixture [**Extended Data Fig. 1B**] [(Yoshihara et al., 2013), and by RNA quality (i.e. detected splice junctions [**Extended Data Fig. 1A]**) were not significantly associated with recurrence. Neither group of samples showed elevated levels of ER (*ESR1*), PR (*PGR*), or HER2 (*ERBB2*) (**Fig. 1D**). Importantly, recurrence was not merely associated with cell proliferation as determined by immunohistochemistry (**Supp. Table 1**) and expression profiling (**Extended Data Fig. 1B-C**), motivating more in-depth analyses.

Association of Tumor Molecular Characteristics with Recurrence

To elucidate the molecular underpinnings of disease recurrence, we probed for differential gene expression followed by functional analyses of the significant genes (**Fig. 2**). First, in a supervised approach, we contrasted recurrent vs. non-recurrent tumors (**Fig. 2A**) and found 560 differentially expressed genes (**Supp. Table 2**) ($p<0.05$, absolute log(fold-change)>0.5). Overall, similar numbers of genes associated with good prognosis (*survival-genes*, n=239), and poor prognosis (*risk-genes*, n=321), where identified. However, slightly more *survival-genes* than *risk-genes* (118 vs. 79) met stringent statistical thresholds ($p<0.01$). Among the most significant risk-genes were *FGF1* and *SLC27A6*. Notably, both genes have a highly restricted expression profile that does not include normal mammary tissues (**Extended Data Fig. 2A,B**). *SLC27A6* is a long-chain fatty acid transporter expressed in mammary epithelial cells. Fatty acid uptake from the surrounding adipose-rich breast tissue is critical for MYC-overexpressing TNBC, which rely on fatty acid oxidation as a source of energy (Camarda et al., 2016). FGF1 is a universal ligand for fibroblast growth factor receptors (FGFRs), and a strong proangiogenic factor (Mori et al., 2013). Paracrine and autocrine FGFR signaling is also a targetable axis (Babina and Turner, 2017) in TNBC due to recurrent genetic aberrations in FGFRs (Meyer et al., 2013; Turner et al., 2010; Wu et al., 2013). Among the most significant survival-genes were *IKZF3* (*Aiolos*) and *FKBP5*, whose expressions are typically restricted to leukocytes (Wang et al., 1998) and adipocytes (Pereira et al., 2014), respectively. These and several other examples in **Supp. Table 2** led us to hypothesize that risk-genes are expressed by cancer cells, while survival-genes arise from constitutive (*e.g.* adipocytes) or infiltrating (*e.g.* leukocytes) cells in the breast tumor microenvironment. Gene-set enrichment analysis supported this observation (**Fig. 2B**). Risk-genes were significantly enriched in epithelial markers (NES=1.72, p=0.0012) and their junctions (NES=1.71, p=0.0025), among several signatures of cellular reprogramming. Survival-genes were strikingly enriched for immune cell markers and immune-related signatures such as lymphocyte activation (NES=-2.71, p=0.0054).

Since BLBC is a heterogeneous disease, we assessed whether recurrence was associated with distinct molecular or histochemical subtypes. We selected the top 200 differential risk- and survival-genes and clustered patients based on their expression (**Fig. 2C**). This supervised assessment revealed three well-defined clusters, each accounting for 24%/48%/28% cases, respectively. Strikingly, recurrence-risk varied considerably between those strata, which we termed low/intermediate/high risk (p=3.9e-6, Fisher's test, two-sided), and ranged between 14% and 74% (**Extended Data Fig. 3A).** The low-risk stratum had high lymphocyte infiltration as assessed by pathologic review (odds-ratio=6.50, p=4.0e-5 by Fisher's test, same below), and showed enrichment for a proposed immunological subtype of TNBC (odds-ratio=21.80, p=1.1e-9) (Lehmann et al., 2016; Ring et al., 2016)). Conversely, the high-risk stratum had low immune infiltration (odds-ratio=6.43, p=2.0e-5), and displayed mesenchymal features (odds-ratio=61.9, p=2.3e-15). When contrasted, high and low risk groups showed pronounced expression differences (**Extended**

Data Fig. 3B). The high-risk stratum was enriched for stem-like, mesenchymal signatures and growth factor signaling, while the low-risk stratum was predictably dominated by immune-related expression (**Extended Data Fig. 3C**). In a parallel unsupervised analysis, we identified independent components (IC) that explain a large proportion of expression variability across patients (Lee, 1998). We found that the first two components were significantly associated with recurrence risk (Methods) (**Fig. 2D**). The second IC subsumed immunological (negative loading) and mesenchymal (positive loading) gene-sets (**Extended Data Fig. 3D**) and was associated with prognosis (**Extended Data Fig. 3E**). Together, these data indicate that recurrence risk in BLBC is significantly correlated with transcriptomic phenotypes that integrate cancer-cell intrinsic and immune-cell extrinsic expression patterns.

Immunogenomic Correlates of Long-Term Recurrence-Free Survival

These results prompted us to explore the immunological differences between recurrent and non-recurrent BLBC tumors in more detail. First, we applied a computational approach, CIBERSORT (Newman et al., 2015), to characterize the tumor micro-environment in terms of magnitude and cellular composition of infiltrating leukocytes. As expected (Jiang et al., 2019; Lehmann et al., 2016), we found pronounced immunological heterogeneity across our cohort (**Fig. 3A**, **Extended Data Fig. 4A-E**). The tumors varied considerably in terms of total levels of immune infiltration as well as compositional characteristics (**Fig. 3A,B**). As expected, tumors stratified by recurrence risk, displayed significant stage-dependent differences in overall immune infiltration ($p<2.2e-16$, Kruskal-Wallis test) (**Extended Data Fig. 4A**). At the most general level, macrophage polarization (M0/M1/M2) was mutually exclusive across patients (**Extended Data Fig. 4B**). High-risk tumors had low macrophage levels and were uncommitted to M0 or polarized towards the immunosuppressive M2 subtype (Hollmén et al., 2015; Sousa et al., 2015) (**Fig. 3C**, **Extended Data Fig. 4C**). Conversely, low-risk tumors had a high proportion of pro-inflammatory M1 macrophages and increased estimated levels of tumor-infiltrating T and B lymphocytes (TILs). RNA-seq enables multiple approaches to quantify tumor infiltration by T and B cells: marker expression, repertoire assembly (i.e. sequencing of the CDR3 complementarity-determining region), and computational cell-type deconvolution. We observed stage-dependent differences in T-cell levels (**Fig. 3D**), with both methods in agreement (**Extended Data Fig. 4D**). Strikingly, we found an equally significant trend for B-cells (**Fig. 3E**, **Extended Data Fig. 4E**). Overall, low-risk tumors showed higher absolute and relative levels of TILs. These increases were mirrored by a lower ratio of regulatory T-cells (Treg) to CD8+ T-cells (fold-change=0.28, $p < 2.2e-16$ by Wilcoxon test, **Fig. 3F**), and higher abundance of Natural Killer (NK) cells (fold-change=1.81, $p=0.0025$ by Wilcoxon test), both indicative of an active anti-tumor immune response in non-recurrent tumors (**Fig. 3G**).

We sought to determine whether the observed immune responses are likely tumor specific. If inflammation is antigen-directed, we expect expansion of select T or B-cell clones (Wells et al., 1997).

We used the Gini index as a statistical measure of inequality of the clonotype distribution (Kirsch et al., 2015). We observed a significant association between the abundance of TILs and the degree of clonal expansion (lower Gini index) (**Fig. 3H**), with a particularly strong correlation (rho=0.69; p<2.2e-16) for B-cells. To explore the novel association between B-cell clonal expansion and cancer recurrence, we constructed a logistic-regression model with both T and B-cell Gini indices as independent covariates (**Supp. Table 3**). Only B-cell clonality remained a significant predictor of recurrence in BLBC (p=0.012). Consistently, among the most significant *survival-genes* (**Fig. 2A**), several were markers of B-cells, including *IKZF3* (fold-change=2.14, p=7.1e-6), *AIM2* (fold-change=2.23, p=0.0016), and *SP140* (fold-change=1.56, p=0.0042). Further, we noted that expression levels of *IKZF3*, a key regulator of B-cell activation (Wang et al., 1998), had a striking dose-dependent association with the likelihood of recurrence (**Extended Data Fig. 4F**) and significantly correlated with activated states of T-cells, B-cells, macrophages, and total immune infiltration score (**Extended Data Fig. 4G**). Taken together, these data indicate that antigen-directed anti-tumor immune responses are strongly prognostic, which prompts the development of immune-focused assays for risk stratification in TNBC.

Development and Validation of Robust Biomarkers for Recurrence-Free Survival in BLBC and TNBC

Given the overall dearth of biomarkers, prognostic stratification remains a challenge in the management of BLBC and TNBC. While expression-based panels, such as MammaPrint (Drukker et al., 2013) and Oncotype DX (Cobleigh et al., 2005) have transformed clinical decision making for ER+ breast tumors, their performance in TNBC/BLBC is poor. In the BRAVO BLBC cohort, MammaPrint- and Oncotype DX-based classifiers showed poor prognostic performance (**Supp. Table 4**), as shown by the lack of differential expression of their constituent genes between recurrent and non-recurrent tumors (**Extended Data Fig. 5A**). Further, with the exception of mesenchymal stem-like tumors, existing molecular subtypes of TNBC are only weakly associated with recurrence (**Fig. 2C**). In our cohort, clinical covariates were also unable to independently predict disease recurrence (**Supp. Table 5**). Informed by the pronounced phenotypic and prognostic stratification of BLBC tumors, we set out to develop a robust panel of prognostic genes and an associated classification algorithm to predict cancer recurrence (**Fig. 4**).

To avoid common pitfalls in statistical biomarker discovery, our approach took advantage of a balanced cohort design (**Fig. 1A**), and applied stringent variable selection and cross-validation (CV) (**Fig. 4A**). First, we noted that *risk-genes* and *survival-genes* were highly correlated and hence of limited utility as independent markers within a prognostic panel (**Extended Data Fig. 5B**). To identify non-redundant and uncorrelated marker genes, we used an ensemble of variable selection algorithms, including information gain (Dai and Xu, 2013) and classification importance (Liaw and Wiener, 2002) (Methods). Their application resulted in a ranking of genes based on both their non-redundancy and expected utility in

classification (**Supp. Table 6**). To determine a useful panel size, we trained a sequence of random forest (RF) classifiers for increasing number of genes with leave-one out CV (LOOCV) (**Extended Data Fig. 6A**). We noticed that average classifier performance started to decline for panels larger than 25 genes, likely due to over-fitting. We selected a 21-gene panel (BRAVO-DX) for further evaluation, as it generalized across different classification algorithms (**Supp Table 7**). Compared to differentially expressed genes, BRAVO-DX markers were, as expected, less redundant and were representatives for larger independent sets of genes with correlated expression patterns (**Fig. 4B**). Notably, BRAVO-DX genes were differentially expressed between recurrent and non-recurrent tumors but did not correlate with histochemical immune responses (**Extended Data Fig. 6C**). Therefore, to assess whether our panel covered both cancer-cell intrinsic and immune-cell extrinsic expression patterns, we correlated the expression of BRAVO-DX genes with predicted abundances of immune cells (**Fig. 4C**). Approximately half of the genes showed negative correlation with the predicted abundance of immune cells and were not immune-cell markers, thus indicating that they were expressed intrinsically by the tumor cells.

Standard classification algorithms optimize classification accuracy by giving equal weight to false-positive and false-negative prediction errors. However, in clinical practice, the sensitivity of picking out cases that are likely to recur (low false-negative) is of utmost importance given the desire to optimize the identification of patients who will develop a recurrence and thus have an appreciably elevated risk of death. Therefore, we used cost-sensitive learning (Elkan, 2001; Schaefer et al., n.d.) to train classifiers with the desired classification characteristic (**Fig. 4D,E**). The baseline RF classifier achieved an overall classification accuracy of 80.7% (15-fold CV) that was balanced in terms of false-positive and false-negative errors. To minimize the number of false-negatives, we systematically increased the ratio of recurrent to non-recurrent cases during training (**Fig. 4D**). As expected, this resulted in a lower number of incorrect negative calls and higher number of false positives. At the optimal oversampling rate of 2.4, we achieved an 84.8% sensitivity in identifying recurrent cases with only a slight decrease in overall accuracy (79.3%) (**Fig. 4E**). Importantly, neither sensitivity nor specificity could be further improved by increasing the panel size or incorporating additional clinical covariates (**Extended Data Fig. 6C**), or using more sophisticated machine learning algorithms such as xgBoost (**Supp. Table 8**). In summary, the balanced design and long-term follow-up of the BRAVO cohort, allowed us to dichotomize recurrence status and employ statistical learning to identify prognostic genes.

To further demonstrate the clinical utility of the BRAVO panel, we proceeded to test its prognostic performance in independent validation cohorts. In addition to the BRAVO-DX 21-gene set, we defined BRAVO-IMMUNE, a 12-gene signature focused on the tumor immune-phenotype (**Supp. Table 7**), and a minimal panel of 3 highly significant risk- and survival-genes *IKZF3*, *AIM2*, and *ELF3*. For evaluation, we selected three large primary breast cancer cohorts with long term follow-up (**Supp. Table 9**). The

three cohorts utilized different transcriptome profiling platforms (i.e., polyA+ RNA-seq, Affymetrix microarray, and Illumina BeadChip) and hence were processed, normalized, and evaluated separately. All patients classified with basal-like or triple-negative diseases were included with no additional inclusion/exclusion criteria. Despite differences in mean follow-up time and event frequency, BRAVO-DX and BRAVO-IMMUNE were highly prognostic in all validation cohorts (**Fig. 5A, Extended Data Fig. 7A**). Notably, BRAVO-DX was more strongly prognostic in patients with favorable prognosis (NPI < 5.4) (**Extended Data Fig. 7B**). The three individual genes, *IKZF3*, *AIM2*, and *ELF3*, were also significantly ($p<0.05$) associated with recurrence-free survival in 8 out of 9 validations (**Fig. 5B, Extended Data Fig. 7C**). Overall, BRAVO signatures had the highest predictive performance for the Gyorffy et al. microarray-based cohort (GEO) (BRAVO-DX: HR=3.45, $p<0.0001$; BRAVO-IMMUNE: HR=4.33, $p<0.0001$), which highlights their ability to generalize across transcriptomic platforms. In order to establish the relative performance of BRAVO-DX, we proceeded to compare it to state-of-art prognostic signatures. We retrieved a battery of 14 marker panels proposed by Rody et al. (Rody et al., 2011), and evaluated them side-by-side with BRAVO-DX and BRAVO-IMMUNE on the BRAVO cohort and 2 validation cohorts (**Fig. 5B, Extended Data Fig. 7D)**. BRAVO signatures achieved highest average accuracy in the discovery cohort (**Fig. 5B**), followed by IFN and MHC-2. In the validation cohorts, BRAVO signatures ranked best in terms of prognostic significance, followed by apocrine and MHC-I panels. Notably, a proliferation-based score was highly significant in the treatment-naive TCGA tumors, but was not validated in the other cohorts. To better understand the functional relationships among signatures, we looked at their correlation within the BRAVO cohort (**Extended Data Fig. 7E**). As expected, BRAVO-IMMUNE was similar to other signatures probing the tumor immune-phenotype. Interestingly, the remaining BRAVO-DX genes correlated most strongly with a signature of claudin-low tumors (Kardos et al., 2016). In summary, we have developed BRAVO-DX, a 21-gene marker panel to distinguish recurrent from non-recurrent BLBC. A BRAVO-DX based RF classifier has demonstrated excellent performance in identifying patients at risk of recurrence in our discovery cohort.

Discussion

We developed and validated a 21-gene signature that is strongly predictive of recurrence free survival in patients with basal-like BCa. As demonstrated by their routine use in the clinic, similar prognostic tools have been highly informative for patients with estrogen receptor positive and HER2-negative (ER+/HER2-) breast cancer. For example, the Oncotype DX test is based on a 21-gene panel and generates a recurrence score used to stratify patients into a high-risk or low-risk group. For high-risk patients, adjuvant chemotherapy is recommended, while the low-risk group can safely avoid chemotherapy and will have an excellent prognosis with endocrine therapy alone (Gluz et al., 2016; Sparano et al., 2018, 2015). Other gene expression profiles, such as PAM50 (Parker et al., 2009), EndoPredict (Dubsky et al., 2013), Breast Cancer Index (Sgroi et al., 2013) and Mammaprint (Drukker et

al., 2013), have also been demonstrated to discern good vs. poor prognosis for ER+/HER2- patients (Harris et al., 2016), and are endorsed for clinical use by the American Society of Clinical Oncology (Harris et al., 2016; Krop et al., 2017). However, there is no analogous test specific for patients with TNBC or BLBC. In contrast to ER+/HER2- disease, adjuvant chemotherapy is the standard of care for all TNBC patients, and hence, the potential clinical use for a prognostic test is different. In particular, a subset of TNBC patients fail standard chemotherapy regimens within the 2-3 years after initial diagnosis. There is, therefore, an unmet clinical need to identify patients who will respond well on standard chemotherapy from those who may have greater clinical benefit from alternate forms of first line therapy.

The molecular characteristics of the BLBC patients and their tumor microenvironment studied here offer insights into alternative first line therapies that may benefit patients who are predicted to rapidly fail standard chemotherapy. Specifically, we observed that the antigen-directed anti-tumor immune response was strongly prognostic. Therapeutic strategies directed toward immunologically “cold” tumors could thus be beneficial for patients identified to have a high risk of recurrence. Alternatively, these patients may also benefit from more aggressive first line chemotherapies such as platinum-based regimens. Indeed, motivating the development of clinical trials to evaluate the effect of such alternative approaches is our validation of our BRAVO-DX signature in three independent cohorts and three different gene expression profiling platforms. Such differences are likely to be non-differential as well as may bias comparisons and reproducibility toward the null.

By utilizing a retrospective cohort of BLBC patients with long-term follow-up and accurate clinical annotation, we were able to identify the transcriptional differences between recurrent and non-recurrent tumors and summarize them as a 21-gene signature. This signature was validated across multiple independent datasets and outperformed existing prognostic markers. Critically, the gene set was robust across transcriptome profiling platforms, which opened up multiple paths to clinical translation. These promising results make BRAVO-DX an excellent candidate for further validations in a prospective-retrospective manner on archival tissues and in prospective clinical trials. In that setting, BRAVO-DX should be compared to pCR and if superior, used within the context of a neoadjuvant therapeutic clinical trial to select patients at high risk of disease recurrence for additional treatment beyond chemotherapy As BRAVO-DX is strongly informed by expression patterns of immune response genes, the signature may ultimately serve as a predictive biomarker to select patients most likely to derive benefit from the addition of immunotherapy-based treatment.

Acknowledgements

This work was supported by the Department of Defense Collaborative Innovators Award to C.I.L. (W81XWH-12-1-0079) and A.M.C. (W81XWH-12-1-0080), National Cancer Institutes to C.I.L. (P50-

CA148143), and Early Detection Research Network (UO1 CA214170) to A.M.C. A.M.C. is a NCI Outstanding Investigator, Howard Hughes Medical Institute Investigator, A. Alfred Taubman Scholar, and American Cancer Society Professor.

Competing Interests

A.M.C. serves on the scientific advisory board (SAB) of Tempus and is a co-founder and SAB member of Lynx Dx. Neither Tempus nor Lynx Dx was involved in the studies presented.

Figures

Figure 1: Study Design and delineation of a basal-like cohort

A) Consort flow diagram of the BRAVO study. Patient eligibility, clinical matching, and cancer recurrence were used to select the BRAVO cohort of Basal-like Breast Cancer (BLBC). **B)** Principal component analysis (PCA) of BLBC in the BRAVO cohort and the breast TCGA cohorts based on genes with most variable expression. **C)** Estimated proportion of tumor cells (tumor purity). **D)** Lack of differential expression of ER (ESR1), PR (PGR), and Her2 (ERBB2), across the BRAVO breast cancer cohorts (paired t-test).

Figure 2: Transcriptional characteristics of recurrent and non-recurrent BLBC tumors

A) Volcano-plot of differentially expressed genes between recurrent and non-recurrent BLBC samples. Genes most significantly associated with recurrence are highlighted. **B)** Gene-set enrichment analysis of molecular signatures from the top up- and down-regulated significant (padj < 0.05) signatures based on the normalized enrichment score (NES) are plotted. In the barcode-plots, genes are plotted left-right from highest to lowest log-fold change. **C)** Supervised clustering of BLBC tumors based on the expression of genes most significantly associated with recurrence. Annotation strips are shown above the expression heatmap: Lymphocyte Response - assessment of tumor-immune infiltration based on pathological evaluation; Subtype - molecular TNBC subtype based on TNBC type; TumorContent - estimated tumor purity (cancer cell fraction). **D)** Unsupervised Independent Component Analysis of variably expressed genes across BLBC. The first two independent components (IC) are shown. Samples are colored according to assignment to risk-group (low/intermediate/high).

Figure 3: Immunogenomic correlates of BCBL recurrence risk

A,B) Tumor infiltration by immune cells is associated with a significant reduction in recurrence risk. **A)** Recurrence risk is highly correlated with magnitude of immune infiltration (CIBERSORT score). **B)** Differences in immune cell-type composition between low/intermediate/high risk groups of BLBC. **C-G)** Association of immune-cell subsets with disease recurrence. **C)** Macrophages and **D)** T cells are based on CDR3 sequence abundance; **E)** B cells are based on CDR3 sequence abundance. **F)** Expression ratio of Treg to all T-cells marker genes (FOXP3 for Treg; CD3 and CD2 for T-cells). **G)** NK-cell abundance based on marker gene expression. **H)** Clonal expansion (Gini index) of T-cells and B-cells is associated with a reduced risk of recurrence.

Figure 4: Development and characterization of a prognostic gene signature

A) Multi-step procedure for the selection of candidate biomarker genes. Genes were pre-filtered based on their expression levels and ranked according to multiple filters that evaluate their independent

classification performance, association with recurrence, and correlation with other genes. Ranking panels of different sizes were then constructed and evaluated. **B)** Correlation heatmap for genes differentially expressed between recurrent and non-recurrent BLBC. BRAVO-DX marker genes select distinct subclusters of correlated genes. **C)** Pearson correlation coefficients of BRAVO-DX marker genes with CIBERSORT estimates of immune cell infiltrates. **D)** Impact of oversampling recurrent cases during training on average random forest classifier accuracy. **E)** Accuracy of an unbiased and cost-sensitive random forest classifier trained to predict cancer recurrence based on the expression of BRAVO-DX biomarker genes.

Figure 5: Validation of the BRAVO-DX gene expression signature

A) Kaplan-Meier plots of progression-free survival in the TCGA and Gyorffy (GEO) cohorts. Patients are optimally dichotomized based on the expression of the complete set of BRAVO-DX marker genes or the BRAVO-IMMUNE subset. **B)** Prognostic performance of multiple gene expression signatures in three independent BLBC/TNBC cohorts. **C)** Prognostic performance of top intrinsic and immune genes within BRAVO-DX signature.

Extended Data Figures

Extended Data Figure 1: RNA-seq Quality Control and Assessment of Confounding Technical Parameters

A) Unsupervised clustering of RNA-seq libraries (columns) based on an array of data-driven quality control measures computed from the raw and aligned sequencing data (Methods). Most informative row-normalized variables are highlighted and correspond to PCR duplication, PCR template switching, frequency of chimeric junctions, and number of detected splice junctions. **B)** Estimated admixture of stromal cells within the tumor tissue based on the expression of signature stromal genes from Yoshihara et. al. **C)** RNA-seq based expression of MKI67 (Ki67), a proliferation marker. RPKM - reads per kilobase per million mapped paired-end reads. Score - Z-score normalized cumulative expression score.

Extended Data Figure 2: Expression of FGF1 and SLC27A6 across normal human tissues

(A) Expression of *FGF1* across the GTEx compendium. **(B)** Expression of *SLC27A6* across the GTEX compendium. *TPM* - transcripts per million.

Extended Data Figure 3: Differential Expression Analysis of High-risk vs Low-risk BLBC tumors

(A) Frequency of recurrence within low/intermediate/high BLBC risk-strata. **B)** Differential expression between high-risk and low-risk BLBC tumors. Highlighted are genes most significantly associated with risk or having the largest absolute fold-changes. **C)** Gene-set enrichment analysis of molecular signatures from genes associated with recurrence risk. The most up- and down-regulated significant signatures based on the negative enrichment score (NES) are plotted (padj < 0.05). Barcode-plots genes are plotted left-right from highest to lowest log-fold change of recurrent vs non-recurrent. **D)** Functional enrichments for genes associated with positive and negative loadings on independent component 2 (IC2). Odds-ratios (solid black line) and significance are shown (right Y-axis); the most informative gene sets are highlighted in red. **E)** Boxplot of IC2 levels across the three BLBC risk-strata.

Extended Data Figure 4: Additional immunogenomic correlates of BCBL recurrence risk

A) Rank-ordered barplot of BLBC samples according to absolute immune infiltration levels (CIBERSORT score). Assignment to BLBC risk-strata is shown below, relative proportions of individual immune cell types are indicated as colored-bands for each sample. **B)** Association of CIBERSORT-estimated immune cell types with recurrence risk groups. CIBERSORT cell-type proportions were averaged across samples within a risk group. **C-G)** Association of predicted immune cell types with recurrence risk: **C)** Macrophages, **D)** T-cells, **E)** B-cells. **F)** Probability of recurrence as a function of IKZF3 expression. **G)** Association (pearson correlation coefficient) of immune-related pathways with IKZF3 expression levels.

Extended Data Figure 5: Patterns of gene expression across the BRAVO cohort.

(A) Expression of OncoType DX and MammaPrint signature genes in the BRAVO cohort. **(B)** Heatmap representing the correlation between genes associated with recurrence (bottom) and survival (top).

Extended Data Figure 6: Optimization of the BRAVO-DX prognostic expression signature

A) Leave-one out cross-validation (LOOCV) of signatures of increasing sizes. Genes are ranked by predictive power and independence. Beginning with MKRN2, genes were progressively added to a growing panel-size. Larger panel-sizes do not result in improved accuracy. **B)** Heatmap representation of the expression of genes in the BRAVO-DX panel. Columns (patients) were divided into recurrent (Rec.) and non-recurrent (Non Rec.) cases. Lymphocyte response was based on pathological assessment. **C)** Impact of clinical covariates on the classification accuracy of a random forest model which added each of the clinical variables as well as BRAVO-DX expression levels.

Extended Data Figure 7: Independent validation of BRAVO-DX

A) Progression-free survival across all TNBC tumors in the METABRIC cohort. Patients are stratified based on the cumulative expression of BRAVO-DX marker genes or BRAVO-IMMUNE (subset of BRAVO-DX that is immune-related). **B)** Same as A) except focused on a subset of patients with the Nottingham Prognostic Index (NPI) < 5.4 (favorable prognosis). **C)** Validation of prognostic utility of three marker genes across the METABRIC cohort. **D)** Associations with survival of multiple prognostic gene-expression signatures across the METABRIC cohort. **E)** Correlations between summary scores of prognostic gene expression signatures in the BRAVO cohort.

Methods

Overall Study Design

To discover and validate a novel prognostic signature specific for BLBC, we utilized a discovery cohort and three validation cohorts. We used the Seattle breast cancer cohort as our discovery cohort because it is highly annotated and includes manually curated medical record data so detailed information on breast cancer recurrences and all primary treatments could be ascertained. Details of this cohort have been previously published (Baglia et al., 2018). Briefly, all participants in this study were identified through the Cancer Surveillance System (CSS), which is the NCI funded population-based Surveillance, Epidemiology, and End Results (SEER) cancer registry serving western Washington state. All study procedures were reviewed and approved by the Fred Hutchinson Cancer Research Center's institutional review board (IRB), and informed consent was obtained from all alive participants. Eligible deceased patients were enrolled through an IRB-approved waiver of consent. Thus, survival bias was reduced in the study as tumor tissue samples were collected and medical records were reviewed for these deceased patients. Data on recurrence status, breast cancer treatments, demographics, lifestyle characteristics, medical history, anthropometric measurements, and family history of cancer were collected through both patient interviews and medical record reviews. >98% of patients enrolled consented to allow us to access their tumor tissue samples, and FFPE samples from consenting participants were requested and attained from the health care providers storing these clinical samples. The three validation cohorts were selected because each included sufficient numbers of TNBC patients with reasonable follow-up time and each had publicly available gene expression data.

Library Preparation and Sequencing

RNA sequencing was performed using standard protocols in our Clinical Laboratory Improvement Amendments (CLIA) compliant sequencing lab (Robinson et al., 2015, 2017). In brief, tumor total RNA was isolated using the AllPrep DNA/RNA/miRNA kit (QIAGEN). RNA sequencing was performed by exome-capture transcriptome platform (Cieslik et al., 2015). All the samples were sequenced on the Illumina HiSeq 2000 or HiSeq 2500 (Illumina Inc) in paired-end mode. The primary base call files were converted into FASTQ sequence files using the bcl2fastq converter tool bcl2fastq-1.8.4 in the CASAVA 1.8 pipeline.

Transcriptomic Data Analysis

Sequence alignment and normalization. Raw sequencing reads were aligned to the GRCh38 reference genome using STAR (Dobin et al., 2013); and overlaps with Gencode v23 (Searle et al., 2010) annotated protein-coding genes were counted using featureCounts (Liao et al., 2013) in strand-specific mode. Non-expressed and lowly expressed genes (<5 reads on average) were removed prior to differential

expression (DE) analysis. A scaling normalization scheme (TMM) was applied to all samples (Robinson and Oshlack, 2010).

Quality control. RNA-seq quality was evaluated by an array of parameters including alignment rate, duplication rate, and number of splice junctions. Samples with multiple parameters falling at the low-quality end were excluded. Samples that passed QC were pooled with TCGA BRCA samples for principal components analysis (PCA). As we used a library preparation method (capture) that was different from TCGA (polyA), we applied adjustment factors to compensate for the systematic discrepancy of the two libraries (Cieslik et al., 2015). PCA was based on the top 500 genes with the largest variance across all samples. PAM50 (Parker et al., 2009) annotation for TCGA dataset was obtained from (Ciriello et al., 2015) and for our samples, and PAM50 was assigned with R package genefu (Gendoo et al., 2016).

Differential analysis. DE analysis was performed with limma (Ritchie et al., 2015) on voom-transformed count data (Law et al., 2014). The top 200 most significant genes (100 genes each for up- and down-regulation) were selected and samples were clustered into 3 groups based on expression of these genes by k-means. Two clusters with high overlap with recurrent and non-recurrent cases were defined as high- and low-risk group, respectively. As the third cluster (intermediate group) exhibited similar gene expression patterns while having an almost equal chance of recurrence and being disease-free, a second round of DE analysis was applied to high- and low-risk groups only to identify stronger signals of recurrence. Gene set enrichment analyses were implemented with R package fgsea (Sergushichev, 2016) on a collection combining gene sets from MsigDB hallmark pathways (Liberzon et al., 2015) and xCell (Aran et al., 2017). R package ica (Helwig, 2018) was used for Independent component analysis (ICA) in order to detect co-expression gene modules.

Immune profiling. Tumor immune infiltration was assessed by multiple computational tools including CIBERSORT (Newman et al., 2015) and MiXCR (Bolotin et al., 2015). CIBERSORT ran with RPKM as input and under “absolute” mode, which returned scores proportional to absolute abundance. By specifying alignment parameters (-p rna-seq), MiXCR took raw mRNA sequences and quantified clonotypes for T- and B-cell receptors.

Biomarker identification. To identify prognostic biomarkers of recurrence, pre-filtered genes (median RPKM>1 and p value<0.1 in differential analysis) were further screened by different feature selection algorithms including Chi2-algorithm (Wang et al., 2005), fast correlation based filter (Yu and Liu, 2003) and information gain (Wang et al., 2005) with R package Biocomb (Novoselova et. al. 2017). Important genes selected by the tree-based classification models random forest (Liaw et al., 2002) and XGBoost (Chen and He, 2015) were also considered as biomarker candidates. All gene candidates were then

ranked by frequency of being selected by the 5 methods as well as by p value from the recurrent versus non-recurrent comparison. This resulted in a panel of 55 genes that were selected by at least 2 methods. We then adopted a stepwise selection strategy to determine the optimal size of the gene panel. Specifically, starting from top 1 gene, gene panels with incremental sizes (adding 1 gene at a time) were evaluated for their ability to correctly classify recurrent and non-recurrent cases by leave-one-out cross validation with random forest. After 21 genes, adding more genes did not benefit overall classification accuracy; therefore the top 21 genes were included in the BRAVO-DX panel. Considering that false-negatives in recurrence predictions are particularly unfavorable,, we used a cost-sensitive variant by oversampling to evaluate the performance of BRAVO-DX with R package mlr in addition to unweighted random forest (Bischl et al., 2016).

Validation Datasets

Three external datasets were used in this study. The first dataset was the BRCA cohort from The Cancer Genome Atlas (TCGA). Raw RNA-seq data were downloaded and processed in the same unbiasedway as in-house sequencing data (except in unstranded mode) to obtain RPKM matrix. The second dataset, GEO, consisted of array-based gene expression profiles compiled from different studies (Györffy et al., 2010). The third dataset was from the METABRIC study (Curtis et al., 2012). Raw idat files were read and processed with BeadArray (Dunning et al., 2007); specifically, normalization across samples were performed with the 'neqc' function (limma) and probes were annotated with illuminaHumanv3.db. Low quality probes (labeled "No match" or "Bad" in the annotation database) were eliminated from further analysis. PAM50 for TCGA dataset was obtained from (Ciriello et al., 2015) and for the other two datasets; R package genefu (Gendoo et al., 2016) was used to assign PAM50.

Survival Analyses

Validation of selected biomarkers was performed on external datasets by univariate survival analysis using the Kaplan-Meier method. Only cases of basal-like subtype, defined by the PAM50 geneset, in both datasets were included. Optimal cutoff was estimated by the cutp function from R package *survMisc*; survival curves were generated with R packages *survival* and *survminer*.

References

Adams, S., Gray, R.J., Demaria, S., Goldstein, L., Perez, E.A., Shulman, L.N., Martino, S., Wang, M., Jones, V.E., Saphner, T.J., Wolff, A.C., Wood, W.C., Davidson, N.E., Sledge, G.W., Sparano, J.A., Badve, S.S., 2014. Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. J. Clin. Oncol. 32, 2959–2966.

Aran, D., Hu, Z., Butte, A.J., 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 18, 220.

Aran, D., Sirota, M., Butte, A.J., 2015. Systematic pan-cancer analysis of tumour purity. Nat. Commun. 6, 8971.

Babina, I.S., Turner, N.C., 2017. Advances and challenges in targeting FGFR signalling in cancer. Nat. Rev. Cancer 17, 318–332.

Baglia, M.L., Cook, L.S., Mei-Tzu, C., Wiggins, C., Hill, D., Porter, P., Li, C.I., 2018. Alcohol, smoking, and risk of Her2-overexpressing and triple-negative breast cancer relative to estrogen receptor-positive breast cancer. International Journal of Cancer.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M., 2016. mlr: Machine learning in R. J. Mach. Learn. Res. 17, 1–5.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Chudakov, D.M., 2015. MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods 12, 380–381.

Brockhoff, G., Seitz, S., Weber, F., Zeman, F., Klinkhammer-Schalke, M., Ortmann, O., Wege, A.K., 2018. The presence of PD-1 positive tumor infiltrating lymphocytes in triple negative breast cancers is associated with a favorable outcome of disease. Oncotarget 9, 6201–6212.

Camarda, R., Zhou, A.Y., Kohnz, R.A., Balakrishnan, S., Mahieu, C., Anderton, B., Eyob, H., Kajimura, S., Tward, A., Krings, G., Nomura, D.K., Goga, A., 2016. Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. Nat. Med. 22, 427–432.

Chen, T., He, T., 2015. Xgboost: extreme gradient boosting. R package version 0. 4-2.

Cieslik, M., Chugh, R., Wu, Y.-M., Wu, M., Brennan, C., Lonigro, R., Su, F., Wang, R., Siddiqui, J., Mehra, R., Cao, X., Lucas, D., Chinnaiyan, A.M., Robinson, D., 2015. The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. Genome Res. 25, 1372–1381.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S.C., Tse, G.M.K., Factor, R.E., Collins, L.C., Allison, K.H., Chen, Y.-Y., Jensen, K., Johnson, N.B., Oesterreich, S., Mills, G.B., Cherniack, A.D., Robertson, G., Benz, C., Sander, C., Laird, P.W., Hoadley, K.A., King, T.A., TCGA Research Network, Perou, C.M., 2015. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell 163, 506–519.

Cobleigh, M.A., Tabesh, B., Bitterman, P., Baker, J., Cronin, M., Liu, M.-L., Borchik, R., Mosquera, J.-M., Walker, M.G., Shak, S., 2005. Tumor Gene Expression and Prognosis in Breast Cancer Patients with 10 or More Positive Lymph Nodes. Clin. Cancer Res. 11, 8623–8631.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Metabric Group, Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J.D., Tavaré, S., Caldas, C., Aparicio, S., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352.

Dai, J., Xu, Q., 2013. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Appl. Soft Comput. 13, 211–221.

Denkert, C., Loibl, S., Noske, A., Roller, M., Müller, B.M., Komor, M., Budczies, J., Darb-Esfahani, S., Kronenwett, R., Hanusch, C., von Törne, C., Weichert, W., Engels, K., Solbach, C., Schrader, I., Dietel, M., von Minckwitz, G., 2010. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J. Clin. Oncol. 28, 105–113.

Denkert, C., Von Minckwitz, G., Brase, J.C., Sinn, B.V., Gade, S., Kronenwett, R., Pfitzner, B.M., Salat, C., Loi, S., Schmitt, W.D., Others, 2014. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2--positive and triple-negative primary breast cancers. J. Clin. Oncol. 33, 983–991.

Dent, R., Trudeau, M., Pritchard, K.I., Hanna, W.M., Kahn, H.K., Sawka, C.A., Lickley, L.A., Rawlinson, E., Sun, P., Narod, S.A., 2007. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin. Cancer Res. 13, 4429–4434.

Dieci, M.V., Criscitiello, C., Goubar, A., Viale, G., Conte, P., Guarneri, V., Ficarra, G., Mathieu, M.C., Delaloge, S., Curigliano, G., Andre, F., 2014. Prognostic value of tumor-infiltrating lymphocytes on residual disease after primary chemotherapy for triple-negative breast cancer: a retrospective multicenter study. Ann. Oncol. 25, 611–618.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Drukker, C.A., Bueno-de-Mesquita, J.M., Retèl, V.P., van Harten, W.H., van Tinteren, H., Wesseling, J., Roumen, R.M.H., Knauer, M., van 't Veer, L.J., Sonke, G.S., Rutgers, E.J.T., van de Vijver, M.J., Linn, S.C., 2013. A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. Int. J. Cancer 133, 929–936.

Dubsky, P., Brase, J.C., Jakesz, R., Rudas, M., Singer, C.F., Greil, R., Dietze, O., Luisser, I., Klug, E., Sedivy, R., Bachner, M., Mayr, D., Schmidt, M., Gehrmann, M.C., Petry, C., Weber, K.E., Fisch, K., Kronenwett, R., Gnant, M., Filipits, M., Austrian Breast and Colorectal Cancer Study Group (ABCSG), 2013. The EndoPredict score provides prognostic information on late distant metastases in ER+/HER2- breast cancer patients. Br. J. Cancer 109, 2959–2964.

Dunning, M.J., Smith, M.L., Ritchie, M.E., Tavaré, S., 2007. beadarray: R classes and methods for Illumina bead-based data. Bioinformatics 23, 2183–2184.

Elkan, C., 2001. The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence. Lawrence Erlbaum Associates Ltd, pp. 973–978.

Gendoo, D.M.A., Ratanasirigulchai, N., Schröder, M.S., Paré, L., Parker, J.S., Prat, A., Haibe-Kains, B., 2016. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics 32, 1097–1099.

Gluz, O., Nitz, U.A., Christgen, M., Kates, R.E., Shak, S., Clemens, M., Kraemer, S., Aktas, B., Kuemmel, S., Reimer, T., Others, 2016. West German Study Group Phase III PlanB Trial: first prospective outcome data for the 21-gene recurrence score assay and concordance of prognostic markers by central and local pathology assessment. J. Clin. Oncol. 34, 2341–2349.

Gonzalez-Angulo, A.M., Timms, K.M., Liu, S., Chen, H., Litton, J.K., Potter, J., Lanchbury, J.S., Stemke-Hale, K., Hennessy, B.T., Arun, B.K., Hortobagyi, G.N., Do, K.-A., Mills, G.B., Meric-Bernstam, F., 2011. Incidence and outcome of BRCA mutations in unselected patients with triple receptor-negative breast cancer. Clin. Cancer Res. 17, 1082–1089.

Győrffy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q., Szallasi, Z., 2010. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res. Treat. 123, 725–731.

Harris, L.N., Ismaila, N., McShane, L.M., Andre, F., Collyar, D.E., Gonzalez-Angulo, A.M., Hammond, E.H., Kuderer, N.M., Liu, M.C., Mennel, R.G., Van Poznak, C., Bast, R.C., Hayes, D.F., 2016. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. J. Clin. Orthod. 34, 1134–1150.

Hollmén, M., Roudnicky, F., Karaman, S., Detmar, M., 2015. Characterization of macrophage--cancer cell crosstalk in estrogen receptor positive and triple-negative breast cancer. Sci. Rep. 5, 9188.

Jiang, Y.-Z., Ma, D., Suo, C., Shi, J., Xue, M., Hu, X., Xiao, Y., Yu, K.-D., Liu, Y.-R., Yu, Y., Zheng, Y., Li, X., Zhang, C., Hu, P., Zhang, J., Hua, Q., Zhang, J., Hou, W., Ren, L., Bao, D., Li, B., Yang, J., Yao, L., Zuo, W.-J., Zhao, S., Gong, Y., Ren, Y.-X., Zhao, Y.-X., Yang, Y.-S., Niu, Z., Cao, Z.-G., Stover, D.G., Verschraegen, C., Kaklamani, V., Daemen, A., Benson, J.R., Takabe, K., Bai, F., Li, D.-Q., Wang, P., Shi, L., Huang, W., Shao, Z.-M., 2019. Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell 0.

Kardos, J., Chai, S., Mose, L.E., Selitsky, S.R., Krishnan, B., Saito, R., Iglesia, M.D., Milowsky, M.I.,

Parker, J.S., Kim, W.Y., Vincent, B.G., 2016. Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. JCI Insight 1, e85902.
Karmakar, S., Harcourt, E.M., Hewings, D.S., Scherer, F., Lovejoy, A.F., Kurtz, D.M., Ehrenschwender, T., Barandun, L.J., Roost, C., Alizadeh, A.A., Kool, E.T., 2015. Organocatalytic removal of formaldehyde adducts from RNA and DNA bases. Nat. Chem. 7, 752–758.
Kawazu, M., Kojima, S., Ueno, T., Totoki, Y., Nakamura, H., Kunita, A., Qu, W., Yoshimura, J., Soda, M., Yasuda, T., Hama, N., Saito-Adachi, M., Sato, K., Kohsaka, S., Sai, E., Ikemura, M., Yamamoto, S., Ogawa, T., Fukayama, M., Tada, K., Seto, Y., Morishita, S., Hazama, S., Shibata, T., Yamashita, Y., Mano, H., 2017. Integrative analysis of genomic alterations in triple-negative breast cancer in association with homologous recombination deficiency. PLoS Genet. 13, e1006853.
Kirsch, I., Vignali, M., Robins, H., 2015. T-cell receptor profiling in cancer. Mol. Oncol. 9, 2063–2070.
Kitano, A., Ono, M., Yoshida, M., Noguchi, E., Shimomura, A., Shimoi, T., Kodaira, M., Yunokawa, M., Yonemori, K., Shimizu, C., Kinoshita, T., Fujiwara, Y., Tsuda, H., Tamura, K., 2017. Tumour-infiltrating lymphocytes are correlated with higher expression levels of PD-1 and PD-L1 in early breast cancer. ESMO Open 2, e000150.
Krop, I., Ismaila, N., Stearns, V., 2017. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology clinical practice focused update guideline summary. J. Oncol. Pract. 13, 763–766.
Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29.
Lee, H.J., Lee, J.-J., Song, I.H., Park, I.A., Kang, J., Yu, J.H., Ahn, J.-H., Gong, G., 2015. Prognostic and predictive value of NanoString-based immune-related gene signatures in a neoadjuvant setting of triple-negative breast cancer: relationship to tumor-infiltrating lymphocytes. Breast Cancer Res. Treat. 151, 619–627.
Lee, T.-W., 1998. Biomedical Applications of ICA. Independent Component Analysis.
Lehmann, B.D., Jovanović, B., Chen, X., Estrada, M.V., Johnson, K.N., Shyr, Y., Moses, H.L., Sanders, M.E., Pietenpol, J.A., 2016. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One 11, e0157368.
Liao, Y., Smyth, G.K., Shi, W., 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics btt656.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news.
Liaw, A., Wiener, M., Others, 2002. Classification and regression by randomForest. R news 2, 18–22.
Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P., 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425.
Lin, N.U., Vanderplas, A., Hughes, M.E., Theriault, R.L., Edge, S.B., Wong, Y.-N., Blayney, D.W., Niland, J.C., Winer, E.P., Weeks, J.C., 2012. Clinicopathologic features, patterns of recurrence, and survival among women with triple-negative breast cancer in the National Comprehensive Cancer Network. Cancer 118, 5463–5472.
Liu, S., Lachapelle, J., Leung, S., Gao, D., Foulkes, W.D., Nielsen, T.O., 2012. CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer. Breast Cancer Res. 14, R48.
Li, X., Krishnamurti, U., Bhattarai, S., Klimov, S., Reid, M., Aneja, R., 2016. Biomarkers Predicting Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer. In: LABORATORY INVESTIGATION. NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA, p. 54A–54A.
Loi, S., Sirtaine, N., Piette, F., Salgado, R., Viale, G., Van Eenoo, F., Rouas, G., Francis, P., Crown, J.P.A., Hitre, E., de Azambuja, E., Quinaux, E., Di Leo, A., Michiels, S., Piccart, M.J., Sotiriou, C., 2013. Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98. J. Clin. Oncol. 31, 860–867.
Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., de Santiago, I., Hopper, J.L., Tsimiklis, H., Apicella, C., Southey, M.C., Schmidt, M.K., Broeks, A., Van 't Veer, L.J., Hogervorst, F.B., Muir, K., Lophatananon, A.,

Stewart-Brown, S., Siriwanarangsan, P., Fasching, P.A., Lux, M.P., Ekici, A.B., Beckmann, M.W., Peto, J., Dos Santos Silva, I., Fletcher, O., Johnson, N., Sawyer, E.J., Tomlinson, I., Kerin, M.J., Miller, N., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Guénel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Bojesen, S.E., Nordestgaard, B.G., Nielsen, S.F., Flyger, H., Milne, R.L., Zamora, M.P., Arias, J.I., Benitez, J., Neuhausen, S., Anton-Culver, H., Ziogas, A., Dur, C.C., Brenner, H., Müller, H., Arndt, V., Stegmaier, C., Meindl, A., Schmutzler, R.K., Engel, C., Ditsch, N., Brauch, H., Brüning, T., Ko, Y.-D., GENICA Network, Nevanlinna, H., Muranen, T.A., Aittomäki, K., Blomqvist, C., Matsuo, K., Ito, H., Iwata, H., Yatabe, Y., Dörk, T., Helbig, S., Bogdanova, N.V., Lindblom, A., Margolin, S., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J.M., Chenevix-Trench, G., kConFab Investigators, Australian Ovarian Cancer Study Group, Wu, A.H., Tseng, C.-C., Van Den Berg, D., Stram, D.O., Lambrechts, D., Thienpont, B., Christiaens, M.-R., Smeets, A., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Radice, P., Peterlongo, P., Bonanni, B., Bernard, L., Couch, F.J., Olson, J.E., Wang, X., Purrington, K., Giles, G.G., Severi, G., Baglietto, L., McLean, C., Haiman, C.A., Henderson, B.E., Schumacher, F., Le Marchand, L., Simard, J., Goldberg, M.S., Labrèche, F., Dumont, M., Teo, S.-H., Yip, C.-H., Phuah, S.-Y., Kristensen, V., Grenaker Alnæs, G., Børresen-Dale, A.-L., Zheng, W., Deming-Halverson, S., Shrubsole, M., Long, J., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Kauppila, S., Andrulis, I.L., Knight, J.A., Glendon, G., Tchatchou, S., Devilee, P., Tollenaar, R.A.E.M., Seynaeve, C.M., García-Closas, M., Figueroa, J., Chanock, S.J., Lissowska, J., Czene, K., Darabi, H., Eriksson, K., Hooning, M.J., Martens, J.W.M., van den Ouweland, A.M.W., van Deurzen, C.H.M., Hall, P., Li, J., Liu, J., Humphreys, K., Shu, X.-O., Lu, W., Gao, Y.-T., Cai, H., Cox, A., Reed, M.W.R., Blot, W., Signorello, L.B., Cai, Q., Pharoah, P.D.P., Ghoussaini, M., Harrington, P., Tyrer, J., Kang, D., Choi, J.-Y., Park, S.K., Noh, D.-Y., Hartman, M., Hui, M., Lim, W.-Y., Buhari, S.A., Hamann, U., Försti, A., Rüdiger, T., Ulmer, H.-U., Jakubowska, A., Lubinski, J., Jaworska, K., Durda, K., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Vachon, C., Slager, S., Fostira, F., Pilarski, R., Shen, C.-Y., Hsiung, C.-N., Wu, P.-E., Hou, M.-F., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M.J., Ponder, B.A.J., Dunning, A.M., Easton, D.F., 2013. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am. J. Hum. Genet. 93, 1046–1060.

Mori, S., Tran, V., Nishikawa, K., Kaneda, T., Hamada, Y., Kawaguchi, N., Fujita, M., Saegusa, J., Takada, Y.K., Matsuura, N., Zhao, M., Takada, Y., 2013. A dominant-negative FGF1 mutant (the R50E mutant) suppresses tumorigenesis and angiogenesis. PLoS One 8, e57927.

Nanda, R., Liu, M.C., Yau, C., Asare, S., Hylton, N., Veer, L.V., Perlmutter, J., Wallace, A.M., Chien, A.J., Forero-Torres, A., Ellis, E., Han, H., Sanders Clark, A., Albain, K.S., Caroline Boughey, J., Elias, A.D., Berry, D.A., Yee, D., DeMichele, A., Esserman, L., 2017. Pembrolizumab plus standard neoadjuvant therapy for high-risk breast cancer (BC): Results from I-SPY 2. J. Clin. Orthod. 35, 506–506.

Network, T.C.G.A., 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A., 2015. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457.

Ono, M., Tsuda, H., Shimizu, C., Yamamoto, S., Shibata, T., Yamamoto, H., Hirata, T., Yonemori, K., Ando, M., Tamura, K., Katsumata, N., Kinoshita, T., Takiguchi, Y., Tanzawa, H., Fujiwara, Y., 2012. Tumor-infiltrating lymphocytes are correlated with response to neoadjuvant chemotherapy in triple-negative breast cancer. Breast Cancer Res. Treat. 132, 793–805.

Paquet, E.R., Hallett, M.T., 2015. Absolute assignment of breast cancer intrinsic molecular subtype. J. Natl. Cancer Inst. 107, 357.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., Bernard, P.S., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167.

Pelekanou, V., Barlow, W.E., von Wahlde, M.-K., Wasserman, B., Lo, Y.-C., Hayes, D.F., Hortobagyi, G.N., Gralow, J., Tripathy, D., Livingston, R.B., Porter, P., Nahleh, Z.A., Rimm, D.L., Pusztai, L.,

2017a. Effects of neoadjuvant chemotherapy (NAC) on tumor infiltrating lymphocytes (TIL) and PD-L1 expression in the SWOG S0800 clinical trial. J. Clin. Orthod. 35, 519–519.
Pelekanou, V., Carvajal-Hausdorf, D.E., Altan, M., Wasserman, B., Carvajal-Hausdorf, C., Wimberly, H., Brown, J., Lannin, D., Pusztai, L., Rimm, D.L., 2017b. Effect of neoadjuvant chemotherapy on tumor-infiltrating lymphocytes and PD-L1 expression in breast cancer and its clinical significance. Breast Cancer Res. 19, 91.
Pereira, M.J., Palming, J., Svensson, M.K., Rizell, M., Dalenbäck, J., Hammar, M., Fall, T., Sidibeh, C.O., Svensson, P.-A., Eriksson, J.W., 2014. FKBP5 expression in human adipose tissue increases following dexamethasone exposure and is associated with insulin resistance. Metabolism 63, 1198–1208.
Perou, C.M., Parker, J.S., Prat, A., Ellis, M.J., Bernard, P.S., 2010. Clinical implementation of the intrinsic subtypes of breast cancer. Lancet Oncol.
Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.-L., Brown, P.O., Botstein, D., 2000. Molecular portraits of human breast tumours. Nature 406, 747–752.
Ring, B.Z., Hout, D.R., Morris, S.W., Lawrence, K., Schweitzer, B.L., Bailey, D.B., Lehmann, B.D., Pietenpol, J.A., Seitz, R.S., 2016. Generation of an algorithm based on minimal gene sets to clinically subtype triple negative breast cancer patients. BMC Cancer 16, 143.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.
Robinson, D.R., Wu, Y.-M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V., Schuetze, S., Alva, A., Siddiqui, J., Chugh, R., Worden, F., Zalupski, M.M., Innis, J., Mody, R.J., Tomlins, S.A., Lucas, D., Baker, L.H., Ramnath, N., Schott, A.F., Hayes, D.F., Vijai, J., Offit, K., Stoffel, E.M., Roberts, J.S., Smith, D.C., Kunju, L.P., Talpaz, M., Cieślik, M., Chinnaiyan, A.M., 2017. Integrative clinical genomics of metastatic cancer. Nature 548, 297–303.
Robinson, D., Van Allen, E.M., Wu, Y.-M., Schultz, N., Lonigro, R.J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C.C., Attard, G., Beltran, H., Abida, W., Bradley, R.K., Vinson, J., Cao, X., Vats, P., Kunju, L.P., Hussain, M., Feng, F.Y., Tomlins, S.A., Cooney, K.A., Smith, D.C., Brennan, C., Siddiqui, J., Mehra, R., Chen, Y., Rathkopf, D.E., Morris, M.J., Solomon, S.B., Durack, J.C., Reuter, V.E., Gopalan, A., Gao, J., Loda, M., Lis, R.T., Bowden, M., Balk, S.P., Gaviola, G., Sougnez, C., Gupta, M., Yu, E.Y., Mostaghel, E.A., Cheng, H.H., Mulcahy, H., True, L.D., Plymate, S.R., Dvinge, H., Ferraldeschi, R., Flohr, P., Miranda, S., Zafeiriou, Z., Tunariu, N., Mateo, J., Perez-Lopez, R., Demichelis, F., Robinson, B.D., Schiffman, M., Nanus, D.M., Tagawa, S.T., Sigaras, A., Eng, K.W., Elemento, O., Sboner, A., Heath, E.I., Scher, H.I., Pienta, K.J., Kantoff, P., de Bono, J.S., Rubin, M.A., Nelson, P.S., Garraway, L.A., Sawyers, C.L., Chinnaiyan, A.M., 2015. Integrative clinical genomics of advanced prostate cancer. Cell 161, 1215–1228.
Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25.
Rody, A., Karn, T., Liedtke, C., Pusztai, L., Ruckhaeberle, E., Hanker, L., Gaetje, R., Solbach, C., Ahr, A., Metzler, D., Schmidt, M., Müller, V., Holtrich, U., Kaufmann, M., 2011. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res. 13, R97.
Schaefer, G., Nakashima, T., Yokota, Y., n.d. Cost-Sensitive Classification for Medical Diagnosis. Encyclopedia of Healthcare Information Systems.
Schmid, P., Adams, S., Rugo, H.S., Schneeweiss, A., Barrios, C.H., Iwata, H., Diéras, V., Hegg, R., Im, S.-A., Shaw Wright, G., Henschel, V., Molinero, L., Chui, S.Y., Funke, R., Husain, A., Winer, E.P., Loi, S., Emens, L.A., IMpassion130 Trial Investigators, 2018. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. N. Engl. J. Med. 379, 2108–2121.
Searle, S., Frankish, A., Bignell, A., Aken, B., Derrien, T., Diekhans, M., Harte, R., Howald, C., Kokocinski, F., Lin, M., Tress, M., Van Baren, M., Barnes, I., Hunt, T., Carvalho-Silva, D., Davidson, C., Donaldson, S., Gilbert, J., Kay, M., Lloyd, D., Loveland, J., Mudge, J., Snow, C., Vamathevan, J., Wilming, L., Brent, M., Gerstein, M., Guigó, R., Kellis, M., Reymond, A., Zadissa,

A., Valencia, A., Harrow, J., Hubbard, T., 2010. The GENCODE human gene set. Genome Biol. 11, 1–1.
Sergushichev, A., 2016. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv.
Sgroi, D.C., Carney, E., Zarrella, E., Steffel, L., Binns, S.N., Finkelstein, D.M., Szymonifka, J., Bhan, A.K., Shepherd, L.E., Zhang, Y., Schnabel, C.A., Erlander, M.G., Ingle, J.N., Porter, P., Muss, H.B., Pritchard, K.I., Tu, D., Rimm, D.L., Goss, P.E., 2013. Prediction of late disease recurrence and extended adjuvant letrozole benefit by the HOXB13/IL17BR biomarker. J. Natl. Cancer Inst. 105, 1036–1042.
Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L.M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S.K., Griffith, M., Moradian, A., Cheng, S.-W.G., Morin, G.B., Watson, P., Gelmon, K., Chia, S., Chin, S.-F., Curtis, C., Rueda, O.M., Pharoah, P.D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J.F., Meyer, I.M., Eaves, C.J., Wasserman, W.W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M.A., Aparicio, S., 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 486, 395–399.
Sigurgeirsson, B., Emanuelsson, O., Lundeberg, J., 2014. Sequencing degraded RNA addressed by 3' tag counting. PLoS One 9, e91851.
Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lønning, P.E., Børresen-Dale, A.L., 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U. S. A. 98, 10869–10874.
Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C.M., Lønning, P.E., Brown, P.O., Børresen-Dale, A.-L., Botstein, D., 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. U. S. A. 100, 8418–8423.
Sousa, S., Brion, R., Lintunen, M., Kronqvist, P., Sandholm, J., Mönkkönen, J., Kellokumpu-Lehtinen, P.-L., Lauttia, S., Tynninen, O., Joensuu, H., Heymann, D., Määttä, J.A., 2015. Human breast cancer cells educate macrophages toward the M2 activation status. Breast Cancer Res. 17, 101.
Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., Geyer, C.E., Jr, Dees, E.C., Goetz, M.P., Olson, J.A., Jr, Lively, T., Badve, S.S., Saphner, T.J., Wagner, L.I., Whelan, T.J., Ellis, M.J., Paik, S., Wood, W.C., Ravdin, P.M., Keane, M.M., Gomez Moreno, H.L., Reddy, P.S., Goggins, T.F., Mayer, I.A., Brufsky, A.M., Toppmeyer, D.L., Kaklamani, V.G., Berenberg, J.L., Abrams, J., Sledge, G.W., Jr, 2018. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N. Engl. J. Med. 379, 111–121.
Sparano, J.A., Gray, R.J., Makower, D.F., Pritchard, K.I., Albain, K.S., Hayes, D.F., Geyer, C.E., Jr, Dees, E.C., Perez, E.A., Olson, J.A., Jr, Zujewski, J., Lively, T., Badve, S.S., Saphner, T.J., Wagner, L.I., Whelan, T.J., Ellis, M.J., Paik, S., Wood, W.C., Ravdin, P., Keane, M.M., Gomez Moreno, H.L., Reddy, P.S., Goggins, T.F., Mayer, I.A., Brufsky, A.M., Toppmeyer, D.L., Kaklamani, V.G., Atkins, J.N., Berenberg, J.L., Sledge, G.W., 2015. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N. Engl. J. Med. 373, 2005–2014.
Trivers, K.F., Lund, M.J., Porter, P.L., Liff, J.M., Flagg, E.W., Coates, R.J., Eley, J.W., 2009. The epidemiology of triple-negative breast cancer, including race. Cancer Causes Control 20, 1071–1082.
Turner, N., Pearson, A., Sharpe, R., Lambros, M., Geyer, F., Lopez-Garcia, M.A., Natrajan, R., Marchio, C., Iorns, E., Mackay, A., Gillett, C., Grigoriadis, A., Tutt, A., Reis-Filho, J.S., Ashworth, A., 2010. FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. Cancer Res. 70, 2085–2094.
Wang, J.H., Avitahl, N., Cariappa, A., Friedrich, C., Ikeda, T., Renold, A., Andrikopoulos, K., Liang, L., Pillai, S., Morgan, B.A., Georgopoulos, K., 1998. Aiolos regulates B cell activation and maturation to effector state. Immunity 9, 543–553.
Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X., Mewes, H.W., 2005. Gene

selection from microarray data for cancer classification—a machine learning approach. Comput. Biol. Chem. 29, 37–46.
Wells, A.D., Gudmundsdottir, H., Turka, L.A., 1997. Following the fate of individual T cells throughout activation and clonal expansion. Signals from T cell receptor and CD28 differentially regulate the induction and duration of a proliferative response. J. Clin. Invest. 100, 3173–3183.
Wu, Y.-M., Su, F., Kalyana-Sundaram, S., Khazanov, N., Ateeq, B., Cao, X., Lonigro, R.J., Vats, P., Wang, R., Lin, S.-F., Cheng, A.-J., Kunju, L.P., Siddiqui, J., Tomlins, S.A., Wyngaard, P., Sadis, S., Roychowdhury, S., Hussain, M.H., Feng, F.Y., Zalupski, M.M., Talpaz, M., Pienta, K.J., Rhodes, D.R., Robinson, D.R., Chinnaiyan, A.M., 2013. Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 3, 636–647.
Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., Carter, S.L., Getz, G., Stemke-Hale, K., Mills, G.B., Verhaak, R.G.W., 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. 4.
Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). pp. 856–863.