AD_____
(Leave blank)

Award Number:
W81XWH-10-1-0299

TITLE:
Uncovering the Hidden Molecular Signatures of Breast Cancer

PRINCIPAL INVESTIGATOR:
Robert Lesurf

CONTRACTING ORGANIZATION:
McGill University, Montreal, Quebec, Canada , H3A 2T5

REPORT DATE:
May 2011

TYPE OF REPORT:
Annual Summary

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 *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
1 May 2011	Annual Summary	1 May 2010 – 30 April 2011

4. TITLE AND SUBTITLE
Uncovering the Hidden Molecular Signatures of Breast Cancer

5a. CONTRACT NUMBER
W81XWH-10-1-0299

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert Lesurf

robert.lesurf@mail.mcgill.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

AND ADDRESS(ES)
McGill University

Montreal, Quebec, Candad H3A 2T5

8. PERFORMING ORGANIZATION REPORT NUMBER

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

It is well understood that breast cancer is a heterogeneous disease, consisting of at least five transcriptional subtypes described by distinct, but poorly understood, molecular profiles. As the cause, aggressiveness, and outcome vary greatly between patients, it is essential to characterize the different ways in which the disease can grow and spread. Transcriptional subtyping operates by capturing the 'loudest' molecular events within a tumor. While these events are both biologically and clinically important, they only represent a fraction of the total cellular pathways and responses. Subtle information is overshadowed by these responses. Such information lies orthogonally to the subtypes, and may be of equal or greater clinical importance. We are proposing a framework to address this challenge. Our methodology is different from what has been done in the past, because it is able to break down tumors using individual signatures. The analysis can be done on a large-scale, and does not require tumors to be binned into distinct classes. In a similar way, murine and cell-line models will be analyzed, allowing us to determine which models best reflect the human disease, and in what way. This will in turn allow us to understand how different tumor processes work together, and to refine our models to better reflect the human disease. We aim to produce an open and accessible framework that will be used to quickly and thoroughly understand the processes that are at play in new tumour cases. This framework will have immediate research applications through the generation of better models for breast cancer. Ultimately, we intend for it

15. SUBJECT TERMS
Breast cancer

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UU	21	**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

Introduction

Breast cancer is a heterogeneous disease, consisting of at least five intrinsic subtypes including the luminal A and luminal B (estrogen receptor alpha positive; ESR+), Her2+ (v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 positive), basal (ESR-, Her2-), and normal-like patient groups[1-3]. These subtypes exhibit distinct differences in their molecular signaling cascades, stress responses, and in the types of cells present within the tumor. For example, the luminal subtypes of breast cancer display a strong estrogen-signaling component, while the Her2+ subtype reflects the downstream response of receptor tyrosine kinase activation. Although such subtype-related pathways and responses are known to be biologically and clinically relevant, they represent only a fraction of total cellular pathways and responses. A more complete understanding is needed to fully determine the reasons for treatment failure and disease recurrence. To date, however, we lack a comprehensive analysis of those processes within the tumor that are associated with outcome (or other histopathological/clinical variables), and whether they are dependent or independent of the tumor subtype. Our previous attempts to answer these questions using traditional bioinformatics approaches appear to fail due to small patient size and the magnitude of the molecular signal involved in different processes and responses.

Our central hypotheses are that each subtype can be defined as a collection of molecular processes, that there exist processes that can be used to predict patient outcome regardless of subtype, and that there exist a disjoint set of processes that predict prognosis within each subtype. Moreover, we argue that the identity of these processes can be inferred through the combined use of our de novo bioinformatics tool entitled Breast Signature Analysis Tool (BreSAT) and our catalogue of transcriptional signatures (entitled BreSAT-DB) that have been collected from literature and resources such as GeneSigDB[4] and MSigDB[5], but carefully modified and augmented to reflect the specific biologies of the breast environment.

Body

Task 1. Complete course requirements (year 1):
1a. BIOC 603: Genomics and Gene Expression (year 1).

All required PhD coursework has been successfully completed. Other program requirements to date, including research seminars 1 & 2 (junior seminar and PhD proposal respectively) were successfully completed.

Task 2. Development of breast cancer-specific signatures (year 1):
2a. Acquire signatures from literature and databases (year 1).
2b. Filter collection based on relevancy (year 1).
2c. Agglomerate signatures representing high biological similarity (year 1).
2d. Refine genes according to behavior in breast-related datasets (year 1).
Milestone #1 Publication (year 1).

A major component of our framework involves the collection and formatting of molecular signatures, along with the development of an appropriate ontological annotation. Signatures are typically a set of genes that have been determined to be differentially perturbed in response to either a specific molecular event (e.g. overexpression of ESR), or are markers of a specific cell type (e.g. macrophages versus pericytes versus endothelial cells). Signature databases such as GeneSigDB[4] and MSigDB[5] exist, and contain thousands of such signatures. However, these signatures have been generated in a variety of organisms, tissues, cell types, and with different techniques. Thus, many of these signatures may not accurately recapitulate the target biology in human clinical breast samples. Furthermore, in some cases, multiple signatures exist for what are meant to be the same biological processes. This creates challenges downstream in the analysis, as separate signatures that represent the same general process or cell type may contain a dissimilar set of genes, which exhibit different expression patterns in human breast cancer data, and ultimately lead to contradictory conclusions. For these reasons, we have refined and annotated thousands of available signatures with features such as the species and tissue they were generated in, as well as their general category (e.g. whether they are used to define a particular cell type, biological response, or a broad prognostic response). Within each of these categories, the signatures are further sub-classified as appropriate (e.g. signatures that define biological responses are sub-classified into one of ten hallmarks of cancer[6]). Our categorizations are intended to allow for the first broad attempt at comprehensively dissecting breast tumors into a set of individual cellular and mechanistic components, and may be further refined and expanded by the community over time.

In addition, we have generated a data compendium now containing over 7000 human patient samples related to breast cancer, along with their associated histopathological/clinical data. Our compendium has been stratified by stages of

disease progression (e.g. normal tissue, DCIS, IDC, metastases, etc.), type of sample (e.g. whole tumor versus cell-specific tissue derived by laser capture microdissection), adjuvant and neoadjuvant treatments, and type of data (e.g. gene expression microarrays, aCGH, miRNA, etc.). While our focus has been on human data, we also have a sizable compendium of models for the disease, including murine tumors and human cell lines. The collection involves a rigorous process of normalization and harmonization. Clinical parameters must be carefully matched to determine, for example, whether recurrence is measured as a local or distant event that takes place in a common 5- or 10-year time frame. This ensures that clinical information is directly comparable from one dataset to the next, and allows us to develop automated tools for analyzing the data.

The collection and annotation of our database and compendium has been relatively straightforward, albeit a time consuming process. A manuscript detailing the production and utility of our databases is currently in preparation, and we expect to submit for publication shortly. After publication, the database and compendium will continue to be updated as new data becomes available.

Task 3. Refinement of statistical methodology (year 1):
3a. Statistic for cohesiveness of subtypes (year 1).
3b. Statistic for association with survival/recurrence (year 1).
3c. Statistic for stability of sample ordering (year 1).

Given a panel of gene expression profiles derived from breast tumor samples, we typically have some information regarding patient clinical attributes including tumor grade, stage, ESR status, Her2 status, lymph node status, and ultimately patient outcome with respect to disease recurrence and overall survival. The canonical example of a question that is asked of such datasets is to identify molecular processes and/or cell types in the tumor that differ between patients of good and poor outcome. It is important to note that the assumption here is that tumors be broadly divided into these two groups before the analysis can be performed. Various bioinformatics tools like GSEA[5,7] exist for this type of analysis. However, the heterogeneity of breast cancer suggests that a simple a priori partition of the patients into classes such as good and bad outcome may not suffice. This is highlighted by the enormous differences that exist between subtypes, and the supposition that tumors of different subtypes recur for separate reasons. Indeed, previous attempts at identifying prognostic predictors of breast cancer outcome have largely been confounded by the subtypes, only having utility in a subset of patients[8]. Our observations suggest that the heterogeneity of breast cancers does not allow such a simple dichotomy, and it is nearly impossible to define 2 or more such classes a priori. Moreover, existing tools such as GSEA have a limitation in that they assume that a process is significantly differentially modulated between the bipartition of the patients. That is, these tools look for sets of genes with high expression in one category but low expression in the other. We argue that it is more natural for samples to display a range of activation levels for a given signature. This

is a biological reality that is accepted within the community, but often ignored by bioinformatics methodologies. For example, it is common for Her2 to be genomically amplified one or more times in breast tumor cells, and its gene expression and membrane protein levels increase continuously in accordance. This increase has been directly linked to a corresponding change in signaling downstream of the receptor[9]. Staining of Her2 by immunohistochemistry (IHC) reveals a continuous range of intensities, which are scored from 0-3+ for simplicity, and often further reduced to simply Her2- or Her2+. While tumors are often summarized by a simple discretization, it is more natural for human breast tumors to display a range in signal activation levels or in the amount of various cell types present; bioinformatics methodologies should reflect this reality.

To overcome this problem, we have designed an intuitive approach that linearly orders tumors over individual signatures (Figure 1), thus measuring the strength of the particular response or cell type within the transcriptional profile of a tumor. Furthermore, in contrast to other traditional methodologies, our approach does not require *a priori* that tumors be binned into distinct classes. As such, the tool allows us to investigate continuous trends across the data, assessing the relative activation of signatures across a panel of patients. Using statistical approaches we have additionally developed, such orderings can be measured for robustness and other assessments of quality.

Since thousands of signatures are being employed, and each one generates a unique patient ordering, we have further developed statistical tests to identify those signatures from this large set that display 'interesting' behavior. The definition of 'interesting' is largely dependent on the particular question being asked of the patient dataset. For example, given a transcriptional signature of ESR activation (that is, the gene set corresponding to transcripts that are differentially expressed when ESR is over-expressed), patients are ordered according to their increasing relative expression of the signature. We may then ask whether the patient order is consistent with other assays for assessing the degree of ESR activity, including for instance IHC staining of the ESR protein (Figure 1). Alternatively, a signature may order patients in such a way that associations can be made with a variety of other histopathological/clinical parameters, such as tumor subtype or patient outcome. The development of statistics to identify such associations has not trivial. For example, in determining an association with patient outcome, the tumor ranks could be treated as a continuous variable under Cox regression, essentially asking whether an increase in patient rank linearly corresponds to a change in patient outcome. Alternatively, the patients on either end of the ordering may share good prognosis, with the patients in the centre of the ordering having poor outcome. Both scenarios present relevant information about how a process or cell type relates to patient prognosis, but they require different means of analysis. There are benefits and drawbacks to the various approaches, and ultimately, any biological conclusions depend on such choices.

We have successfully developed a variety of statistics that are able to determine associations between the patient ordering and discrete clinical variables (such as ESR status or tumor subtype), continuous variables (such as age), as well as patient outcome. In addition, we have developed statistics that measure the stability of a patient ordering generated by a particular signature, when compared against the stability generated by a random set of genes. This allows us to filter out those signatures that are less trustworthy in the data.

The type of statistic described thus far treats each signature independently. However, a natural question arises as to whether dependencies exist between the patient orderings generated by each signature. There may be technical reasons for dependencies between signatures (e.g. they have many genes in common), or there may be some underlying biological reason. For such a set of signatures that order patients in a similar way, we wish to investigate whether they also tend to share associations with histological/clinical parameters and/or functional ontologies. To investigate this, we begin by calculating the correlation between every pair of patient orderings, and use this information to build a graph network with edges placed between nodes (signatures) that have a high correlation (figure 2). Highly interconnected regions of the graph are investigated for overrepresentations in associations with available histological/clinical parameters. This is not simply a technical investigation, but one with biological and clinical worth. The fact that processes are correlated tells us about how tumor cells respond to stress, and hints at the molecular level regulatory interactions that take place in tumor progression. This in turn suggests better stratification of patients for the development and success of new treatment targets. Thus, such a signature-network approach identifies functionally-related signatures, even when the signatures represent different biological processes that share little or no genes in common. However, the precise means of building and subdividing such a network is not trivial, from either a computational or an analytical view. Finding the exact means of doing so is an ongoing process.

Task 4. Application of framework to datasets (year 1-3):
4a. Apply signatures to human tumor datasets (year 1-2).
Milestone #2 Publication (year 2).

Our linear ordering procedure has been repeated for ~2000 signatures within our catalogue BreSAT-DB, across a compendium of ~1800 invasive breast carcinomas for which clinically annotated whole tumor gene expression data was available. Appropriate tests were used to identify statistically significant associations between the patient ordering generated by each signature, and histopathological/clinical variables including intrinsic subtype, ESR status, Her2 status, lymph node status, grade, recurrence, and overall survival. An interesting early finding was that the large majority of signatures have a significant association with certain clinical variables, such as ESR status and the tumor subtype. In fact, even random sets of genes tended to produce significant associations. This is a testament to the

enormous transcriptional perturbations that occur downstream of specific molecular events, including activation of ESR. To compensate for this trend, the significance of an association with a given molecular signature is adjusted by resampling 10,000 random gene sets of the same size.

After adjustment, there remained a large number of signatures consistently having significant associations with the variables tested, and there was a surprising overlap in the signatures that associate with any given variable. (figure 2,3). 239 signatures consistently had a significant association with molecular subtype in at least half of the datasets investigated (adjusted pvalue <= 0.05). Typically this association was the result of Luminal A and Basal tumors having vastly different patient ranks. In addition, 207 signatures were found to consistently have significant associations with ER status, 23 with lymph node status, 125 with disease recurrence, and 116 with overall survival (161 combined total for patient outcome). As expected, signatures designed to predict patient outcome in breast cancer patients were all highly significant in the majority of datasets. Remarkably, however, we have been able to identify signatures with consistent, significant associations to patient outcome, but having no such associations to any of the other variables tested. These are signatures that encompass a variety of processes, such a response to hypoxia, VEGF signaling, or activation of the complement immune system. Because such signatures operate independently of known histopathological/clinical parameters, they represent a unique class with prognostic value across all subtypes, which contrasts the types of predictors that are in clinical use[8]. This is an important milestone, because it identifies molecular markers that are determinants of outcome in breast cancer, but have remained unrecognized to date. The identification of such elements is essential towards the development of new classes of treatments. Furthermore, our methodology represents a fundamentally different way of characterizing breast tumors. Whereas traditional approaches segment patients into classes according to the expression of a small number of genes, BreSAT comprehensively identifies the entire set of pathways, processes, responses, and cell types that define the disease. This exhaustive cataloguing of the molecular differences between subtypes is providing a more refined understanding, clinically and molecularly, of the underlying biology of the disease.

As there is some indication that breast tumors of each intrinsic subtype represent distinct biological entities, our analysis was further extended to observe how signatures associate with histopathological/clinical variables within each individual subtype. BreSAT was applied in isolation to patient sets belonging to each of the five intrinsic subtypes, and statistical associations were determined as before. Interestingly, these results revealed that each subtype tends to favour its own set of signatures (and by extension, processes) that associate with patient outcome. The luminal A subtype contained the largest number of signatures that were associated with patient outcome (recurrence and/or overall survival), most of which ordered patients in a manner that was independent of ER status, LN status, and grade. In contrast, tumors belonging to the luminal B subtype had only 7 signatures consistently associated with patient outcome in at last half of the datasets tested.

Surprisingly, 5 of these 7 were signatures derived to specifically predict outcome in breast cancer patient. This suggests that patients with luminal B tumors are especially good candidates for therapeutic decision-making through genomic predictors. Tumors within the ERBB2 and Basal subtypes also had a small number of associations between signatures and patient outcome (8 and 2 respectively), possibly due to the smaller sample size of these subtypes. These associations related to processes such as TGF-Beta and p21 in the ERBB2 subtype, and CK1 and mRNA processing in the Basal subtype. The disparities in the results are perhaps not surprising, as the patients with tumors belonging to different subtypes tend to receive different treatments for their disease. However, our results are particularly applicable as indicators of how and why current treatments fail in different subsets of breast cancer patients.

Such results support our hypotheses that breast tumors can be described by the activation/repression of various molecular signatures, which can act in parallel or orthogonally to a tumor's intrinsic subtype, and are a consequence of the complex mix of cell types within the tumor. To better understand the contribution of different cell types to breast tumor biology and disease outcome, we next applied BreSAT to a dataset containing microdissected epithelium and stroma tissue from matched breast tumors (figure 4). As before, statistical tests were used to identify associations between signatures and histopathological/clinical variables of interest. Because the process was performed in matching tumor epithelium and stroma, we were able to distinguish between signatures that are macroenvironmental (present in all compartments of the tumor) vs those that are microenvironmental (present either in epithelium or stroma, but not both). Furthermore, our results have revealed that some subsets of patients display remarkably similar signature activation/repression in matched tumor epithelium and stroma, whereas other patient subsets are enriched in microenvironment-specific responses.

We are additionally in the process of investigating the types of dependencies that exist between signatures. By quantifying the correlation between all possible pairs of signature-derived patient orders, we hope to identify functional associations between signatures, even when the signatures represent vastly different biological processes that share little or no genes in common. Our initial analysis indicates that although there is an overrepresentation of highly correlated signatures with a significant number of genes in common, there additionally exist many correlated signature pairs with no overlap. We identify many such distinct types of processes and cell types that appear to be highly correlated to one-another, and are currently examining ways of subdividing our collection of signatures into a core set of groups. The fact that many processes are co-modulated suggests methods for building more robust and accurate prognostic signatures, that encompass a broader range of clinically-relevant characteristics with highly resilient signals.

A manuscript detailing the application of our framework to the whole tissue datasets is in preparation and on track for submission this year. The results derived

from the laser capture microdissected tumor data will be submitted as a separate publication.

Task 5. Hypothesis-driven generation of model systems (year 2-3):
5a. Selection of appropriate cell lines and mouse models (year 3).
5b. Molecular engineering of models (year 3).
5c. Analysis of modification success (year 3).
Milestone #3 Publication (year 3).

This task largely represents work to be done over the next two years of the project. Several hundred samples have already been collected and formatted for various mouse models and cell lines of the disease. This data collection will continue in the future. Upon completion of earlier tasks, the *de novo* methodologies we have developed will be applied to this data. Comparisons will be made with the results from *in vivo* human samples, in order to better characterize those pathways, processes, responses, and cell types that are shared or differ between models of the disease and the disease itself.

Key Research Accomplishments

- Construction of highly annotated signature database, specific to breast cancer (BreSAT-DB).

- Collection and formatting of over 7000 data samples relating to breast cancer, primarily gene expression profiles of invasive ductal carcinoma (BreSAT-Compendium).

- The generation of various statistical methodologies to apply signatures to the collected datasets, thereby determining the significance of associations between pathways, processes, responses, or cell types, and available histopathological/clinical parameters.

- Application to of our signatures to human datasets, testing for statistical associations and dependencies between signatures.

Reportable Outcomes

Poster
Title: Breast Signature Analysis Tool (BreSAT): a framework for investigating the molecular networks of breast cancer
Conference: RECOMB Computational Cancer Biology 2010
Location: Oslo, Norway
Date: June 2010

Presentation
Title: Breast Signature Analysis Tool (BreSAT): a framework for investigating the molecular networks of breast cancer
Conference: 10th Annual McGill Workshop on Bioinformatics in Barbados: Systems Approaches in Translational Breast Cancer Research
Location: Holetown, Barbados
Date: January 2011

Collection and normalization of breast related data (BreSAT-Compendium)
In total, our compendium now includes over 7000 human patient samples with associated histopathological/clinical data. Our compendium has been stratified by stages of disease progression (e.g. normal tissue, DCIS, IDC, metastases, etc.), type of sample (e.g. whole tumor versus cell-specific tissue derived by laser capture microdissection), adjuvant and neoadjuvant treatments, and type of data (e.g. gene expression microarrays, aCGH, miRNA, etc.). While our focus has been on human data, we also have a sizable compendium of models for the disease, including murine tumors and human cell lines. The collection involves a rigorous process of normalization and harmonization. Clinical parameters must be carefully matched to determine, for example, whether recurrence is measured as a local or distant event that takes place in a common 5- or 10-year time frame. This ensures that clinical information is directly comparable from one dataset to the next, and allows us to develop automated tools for analyzing the data.

Construction of signature database (BreSAT-DB)
Collection, refinement, and annotation of over 6000 available molecular signatures with features such as the species and tissue they were generated in, as well as their general category (e.g. whether they are used to define a particular cell type, biological response, or a broad prognostic response). Within each of these categories, the signatures are further sub-classified as appropriate (e.g. signatures that define biological responses are sub-classified into one of ten hallmarks of cancer[6]).

Conclusion

The framework we have described is a novel and important step towards better understanding the underlying pathways, processes, responses, and cell types that influence breast cancer progression and outcome. Our data compendiums represent the largest effort we are aware of to collect high-throughput breast-related data in an appropriately formatted and clinically annotated fashion. Similarly, our signature collection BreSAT-DB, contains the largest signature collection known to us, and is additionally the most thoroughly annotated and relevant to breast cancer. Work will continue on the projects/manuscripts currently in preparation, with the goal of publishing within the near future. In conjunction with the publications, we aim to release our methodologies as an open source R/bioconductor package.

Our analysis with the BreSAT framework is allowing us to piece together the interplay between individual molecular signatures, and to better understand how this interplay affects the phenotype of breast cancer. BreSAT provides the means to comprehensively identify the pathways, processes, responses, and cell types that impact tumor progression. Our methodology introduces a unique and intuitive semi-supervised approach to pathway analysis, and is robust when multiple disparate high-throughput datasets are used. Crucially, it represents an entirely different way of classifying the disease. Instead of relying on the 'loudest' molecular signals that make up the majority of a transcriptional profile, the status of subtle but important biological pathways are taken into account. BreSAT provides a means to determine the classes of responses that characterize patient outcome, regardless of the confounding effect of the tumor subtype. Together, this work helps provide an essential step forward in understanding the molecular components that are involved in breast cancer.

Our thorough analysis of primary human tumors is inevitably generating new hypotheses about the disease. For future validation, model systems of human breast cancer are needed. One of the goals of our work is to develop a means of mapping mouse and cell-line models to appropriate human tumors. Having broken down the biological activity of the tumors into their core components, we will next seek to identify from these components exactly what is shared between individual mouse or cell-line models and human tumor subtypes. In the same way that BreSAT was applied to primary human tumors, it will also be applied across our compendium of model systems. The behavior of signatures can then be systematically compared, allowing us to determine which models best reflect the human disease, and in what way.

References

1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.

2. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lønning P, Børresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11;98(19):10869-74.

3. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale AL, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003 Jul 8;100(14):8418-23.

4. Culhane AC, Schwarzl T, Sultana R, Picard KC, Picard SC, Lu TH, Franklin KR, French SJ, Papenhausen G, Correll M, Quackenbush J. GeneSigDB--a curated database of gene expression signatures. Nucleic Acids Res. 2010 Jan;38(Database issue):D716-25.

5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646-74.

7. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003 Jul;34(3):267-73.

8. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med. 2006 Aug 10;355(6):560-9.

9. Hodgson JG, Malek T, Bornstein S, Hariono S, Ginzinger DG, Muller WJ, Gray JW. Copy number aberrations in mouse breast tumors reveal loci and genes important in tumorigenic receptor tyrosine kinase signaling. Cancer Res. 2005 Nov 1;65(21):9695-704.

10. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 Dec 19;347(25):1999-2009.

11. Abba MC, Hu Y, Sun H, Drake JA, Gaddis S, Baggerly K, Sahin A, Aldaz CM. Gene expression signature of estrogen receptor alpha status in breast cancer. BMC Genomics. 2005 Mar 11;6:37.

12. Finak G, Sadekova S, Pepin F, Hallett M, Meterissian S, Halwani F, Khetani K, Souleimanova M, Zabolotny B, Omeroglu A, Park M. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. Breast Cancer Res. 2006;8(5):R58.

13. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30.

Appendices

Not Applicable.

Supporting Data

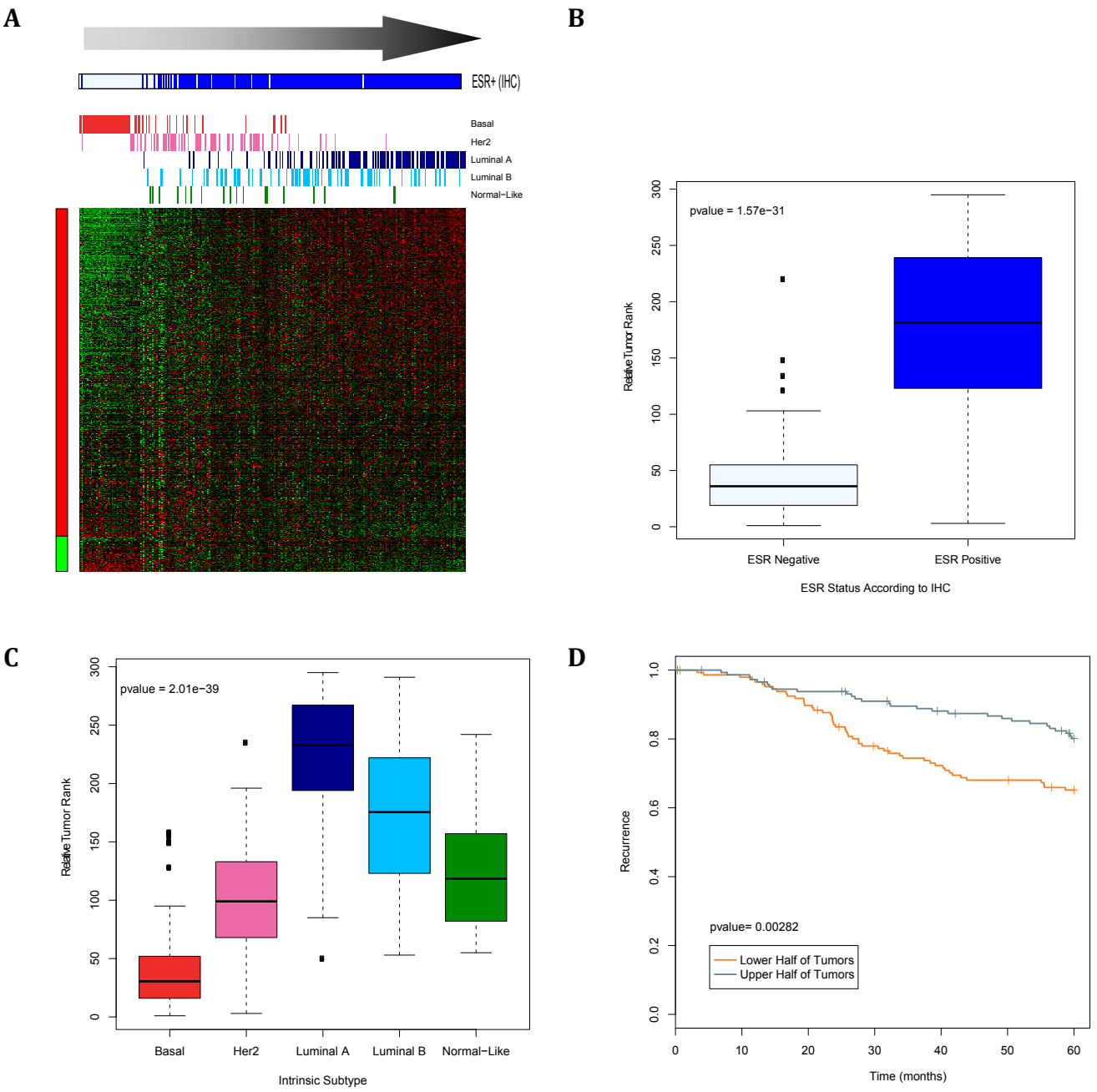

Figure 1. ESR activation signature in a breast cancer dataset. (A) Heatmap of ESR activation signature, with rows representing genes, and columns representing tumors. Gene expression is colored from green (low) to red (high). Samples are ordered from left (least ESR signaling activation) to right (most ESR signaling activation) using BreSAT. Arrow indicates increasing signature activation in the tumors. Patients are labeled according to their ESR IHC status (blue=positive), and their intrinsic subtype. (B) Patients ranks of the ESR- and ESR+ classes are displayed as boxplots, and are significantly different (p-value=1.6×10^{-31}). (C) Patient ranks of the intrinsic subtypes are displayed as boxplots, and are significantly different (p-value=2.0×10^{-39}). (D) Tumors were broadly divided in half according to their ranks, and Kaplan-Meier curve shows tumor recurrence of the two groups. The tumors with less ESR signaling activation have significantly worse outcome (p-value=2.8×10^{-3}). Expression data was obtained from [10]; ESR activation signature was obtained from [11].

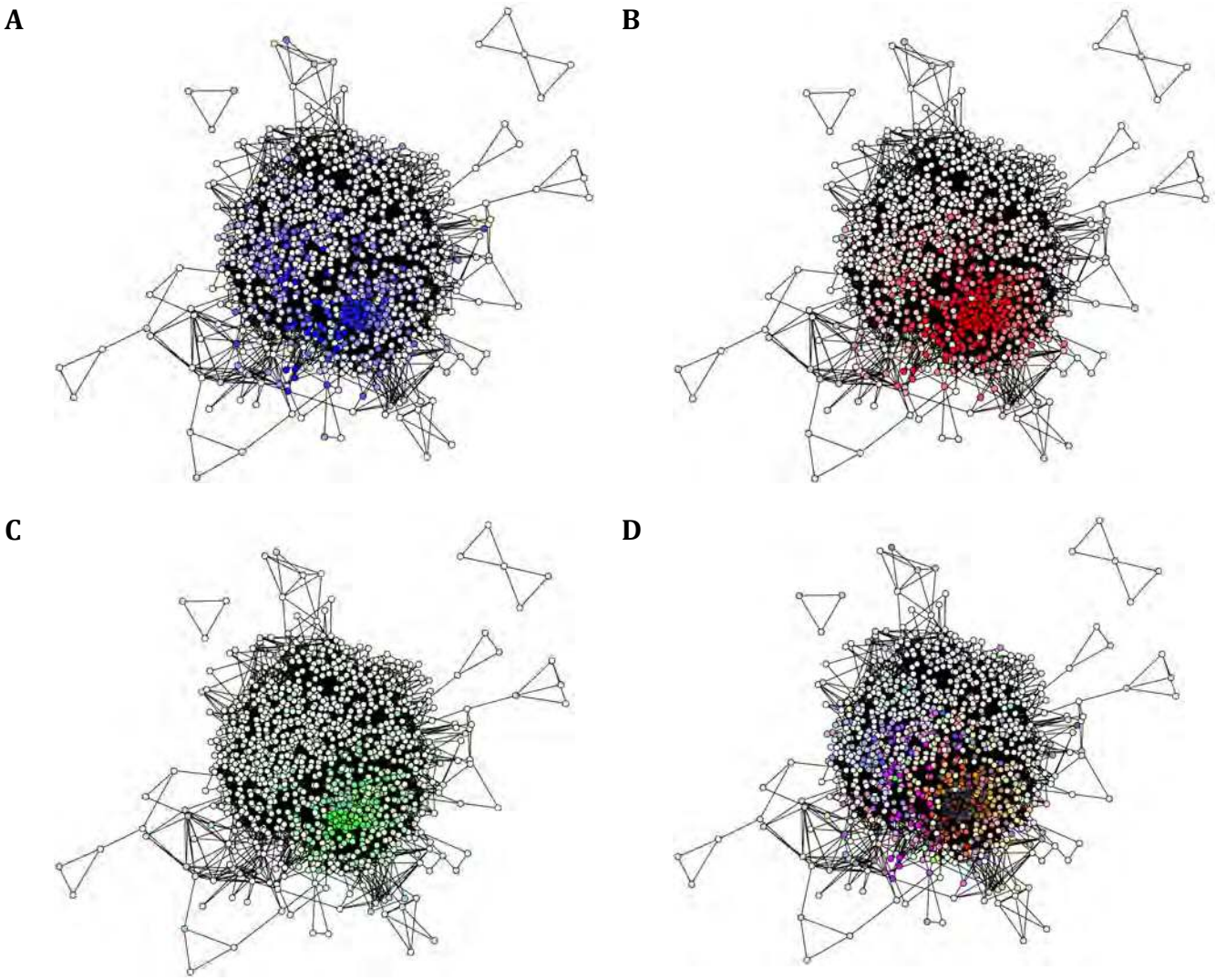

Figure 2. Network view of correlations between signature orderings. Nodes represent each signature tested, and are joined by edges representing the highest positive 1% and negative 1% of median correlations between signature ordering pairs across datasets. Nodes are colored according to the proportion of datasets where they have significant associations with ESR status (A), subtype (B), recurrence (C), and the overlap (D).

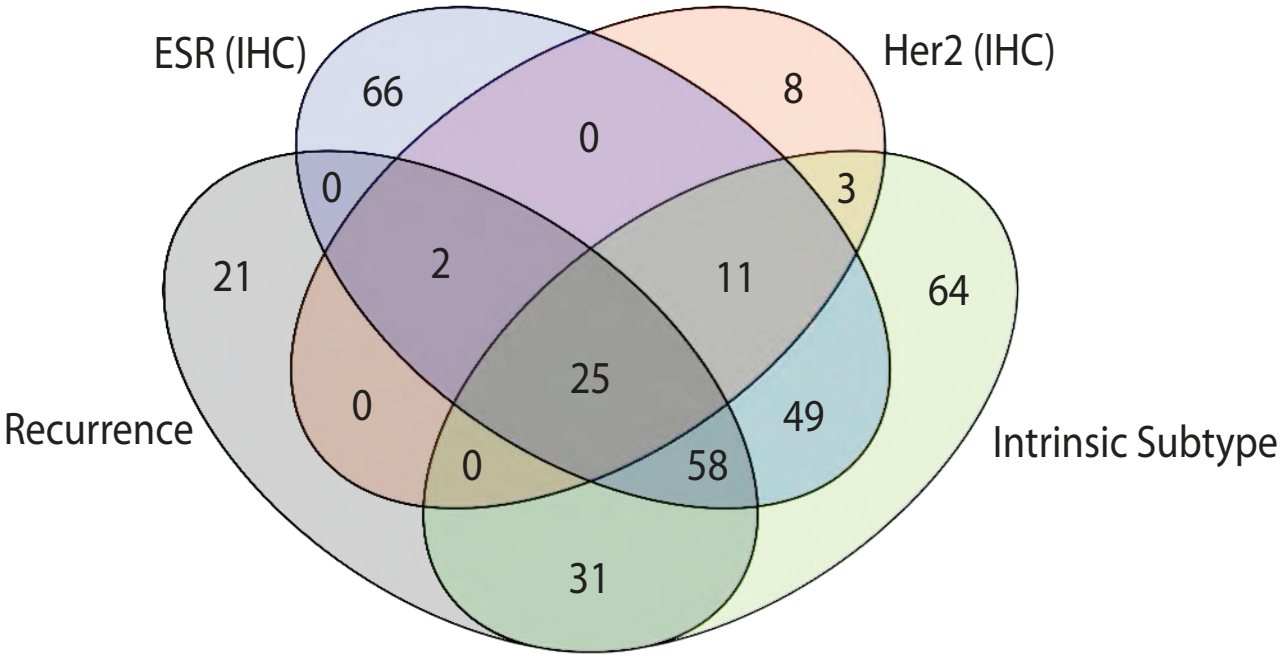

Figure 3. Venn diagram representing significant clinical associations. Signatures must be significantly associated (adjusted p-value<0.05) with ESR status, Her2 status, intrinsic subtype, and/or disease recurrence, in at least 50% of datasets tested. 21 signatures were found to be uniquely associated with recurrence.

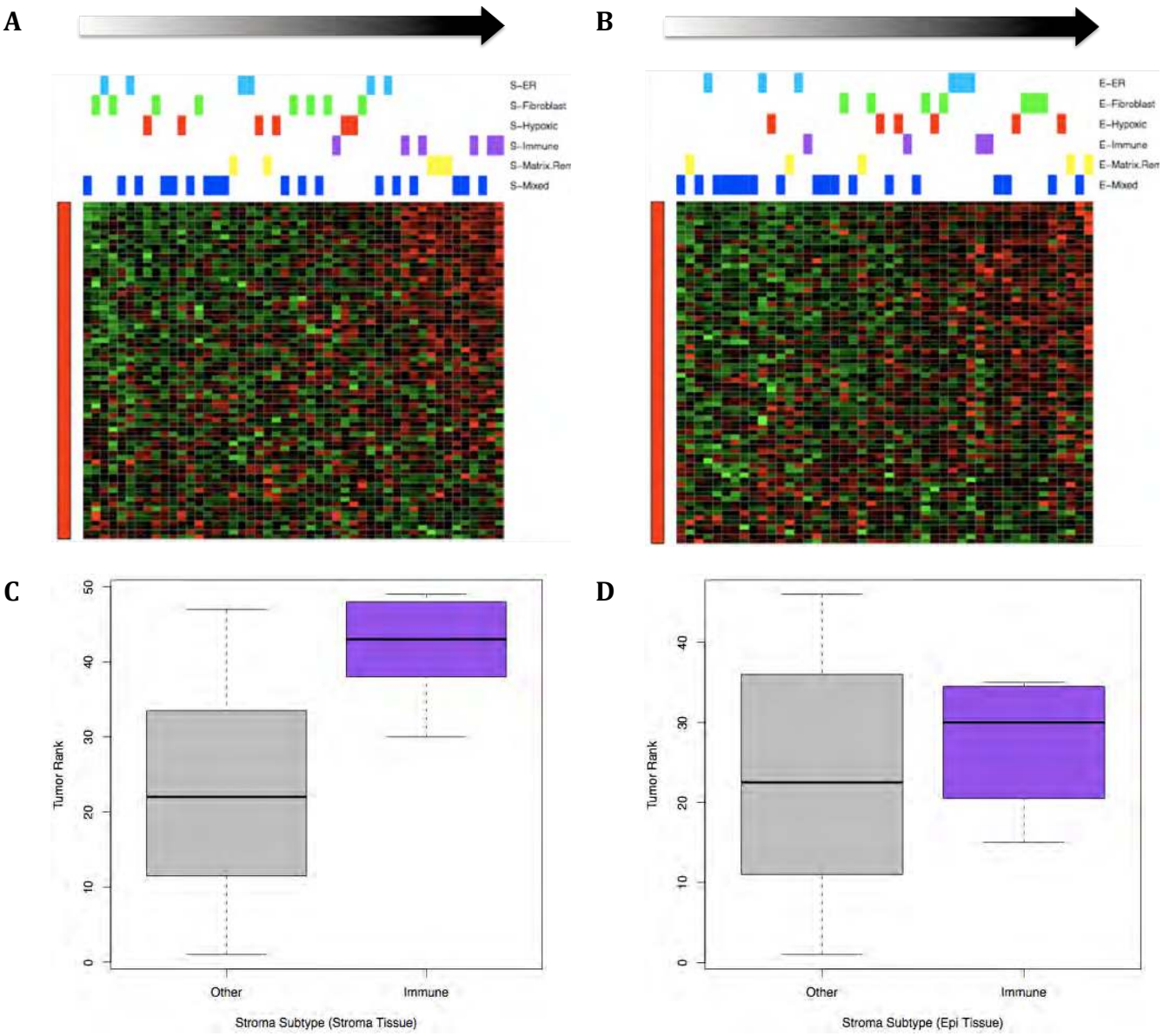

Figure 4. Natural killer cell-mediated cytotoxicity activation signature in stromal and epithelial breast tissue. (A) Heatmap showing laser capture microdissected stromal tissue ordered from left (representing less activation of the signature) to right (representing greater activation of the signature). Samples are labeled according to their intrinsic stromal subtype: ER high (light blue), fibroblast-enriched (green), hypoxic (red), immune-enriched (purple), matrix remodeling (yellow), and mixed (dark blue). (B) Heatmap showing laser capture microdissected epithelial tissue from the same tumors as in A, and labeled according to their intrinsic stromal subtype. Boxplots of the patient rank distributions for immune-enriched (purple) and all other samples (gray) in stromal tissue (C) and epithelial tissue (D). Immune-enriched stromal tissue shows significantly greater activation of the signature (p-vlaue=8.99×10^{-4}), while the epithelial tissue does not (p-value=0.560). Expression data was obtained from [12]; the signature was obtained from [13].