IDENTIFYING KEY GENE SETS IN METASTATIC DORMANCY: ONTOLOGY ENRICHMENT IN GENE EXPRESSION PROFILES FROM SELECTED HUMAN BREAST CANCER PATIENTS

by

Adam B. Pater-Faranda

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Fall 2018

IDENTIFYING KEY GENE SETS IN METASTATIC DORMANCY:

ONTOLOGY ENRICHMENT IN GENE EXPRESSION PROFILES

FROM SELECTED HUMAN BREAST CANCER PATIENTS

by

Adam B. Pater-Faranda

Approved: _____

April M. Kloxin, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Kathleen F. McCoy
Chair of the Department of Computer and Information Sciences

Approved: _____

Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved:

_____

Douglas J. Doren, Ph.D.
Interim Vice Provost for the Office of Graduate and Professional Education

ACKNOWLEDGMENTS

To the members of my committee: Dr. April Kloxin, Dr. Karen Ross, and Dr. Cathy Wu. I want to express my sincere gratitude for your guidance and support. To Dr. Ross and Dr. Kloxin, with whom I have worked so closely these past 18 months, I deeply appreciate our many stimulating conversations during this project and the thought and care you put in to helping me develop the work. Dr. Ross, I thank you especially for the expertise you provided while guiding me through the evaluation of the gene expression and enrichment results presented herein. Thank you all for making this important milestone possible.

Thank you to the members of the Kloxin group who assisted in querying DAVID with randomized lists. Your help expedited the process tremendously.

To my parents Tom and Lisa, my sister Elizabeth, her husband Derek, and my love, Jerilyn Jones. Thank you so much for your encouraging me whenever I felt overwhelmed, and helping me find the determination to continue.

To Maria Donner, and Abby Myhre: Thank you so much for having mentored me and cultivated my development in Genetic Toxicology. You've encouraged my studies since I first took Cancer Biology in 2012; I no longer wait 3 months to check my grades! Thank you so much for your stalwart support of my continued education.

TABLE OF CONTENTS

Appendix

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

For approximately 20% of breast cancer patients, many years pass between initial remission and the emergence of distal metastases. These late recurrences are hypothesized to arise from malignant cells shed from the primary tumor that remain dormant in secondary tissues until resuming metastatic proliferation. A central question is the extent to which the dormancy interval is determined by the initial state of cells in the primary tumor, and whether proliferation resumes as the result of changes in the distal microenvironment. Differential gene expression in primary tumors may provide insight into the processes involved in the maintenance of metastatic dormancy.

Four clinical studies providing microarray profiles were identified based on specific endpoints, and follow-up duration. Patient groups were assembled based on disease status and the time to distant metastases. Differential gene expression and ontology enrichment for these patients was evaluated in an automated workflow.

Despite some weakness in differential expression, noteworthy overlap between studies was observed at the level of ontological enrichment. Several biologically relevant terms were detected in enrichments for three or more studies. These included terms such as "*platelet degranulation*", "*wound healing*" and "*cell proliferation*" all processes that may be are involved in the escape from dormancy.

Chapter 1

BREAST CANCER, DORMANCY, AND GENE EXPRESSION SIGNATURES

Breast Cancer and Estrogen Receptor Status

Breast cancer is a disease of the breast that involves the formation of a tumor from a population of abnormally proliferating cells. At an anatomical level, the majority of breast tumors arise from the epithelium lining the milk ducts. The SEER program is a major cancer surveillance initiative in the United States. Their data for 190,458 women shows that between 1987 and 1999 invasive ductal carcinoma was observed in the vast majority of cases (72.8 %), followed by invasive lobular carcinoma (7.6%) and invasive ductal-lobular carcinoma (4.7%) [1].

Breast tumors can also be characterized based on a variety of molecular biomarkers. Today, three of the most important markers are the estrogen receptor (ER), the progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). It is common for ER and PR status to correlate with one another [2]. The estrogen receptor was one of the first clinically useful markers, identified in the early 1970's. As early as 1974, it was becoming apparent that patients with estrogen receptor (ER) positive tumors responded well to anti-estrogenic therapies while patients with ER negative tumors did not [3]. In a later study, results show significant differences in the site of first metastasis between patients with ER positive and ER negative tumors. Estrogen receptor status can influence the distribution of organs and tissues affected by metastatic disease [4].

The later stages of breast cancer progression often involve the emergence of secondary tumors known as metastases. Metastases originate from malignant cells that dissociate from the primary tumor mass, travel through the body's circulatory system and settle in a distant tissue. Once a tumor cell has arrived at a distal site, it may continue to proliferate forming a new tumor [5,6]. This process is often described as a "Metastatic Cascade" [5,6]. While inherently inefficient, metastasis is the most frequent proximal cause of breast cancer fatality [6].

As early as 1889, mounting evidence indicated that cancers of the breast were predisposed to form metastases in the bone [7]. Since Stephen Paget first proposed his hypothesis of "seed and soil" we have learned a great deal about which tissues are invaded by metastatic breast cancer. In a study of breast cancer patients treated at University of Maryland hospitals, bones were the most frequent site of metastasis, followed by lung, liver and brain [8]. In a larger study, spanning 4399 patients, distal metastases were most frequently observed in lymph nodes, bone, liver and lung [9]. Studies have shown significant differences in the site of first metastasis between patients with ER positive and ER negative tumors. In this study, the site of first metastasis for ER positive tumors was predominantly bone vs predominantly visceral sites such as the liver, lungs, brain and ovaries, for ER negative tumors [4].

Metastatic Dormancy

A defining feature of estrogen receptor (ER) positive breast cancer is the phenomenon of metastatic dormancy. Patients with this disease can experience

metastatic recurrence at distal sites long after their primary breast tumor has been treated [10]. Even for women who had no positive lymph nodes when treated the risk of recurrence after 20 years may be as high as 15%, and greater for those with more significant node involvement [10]. These late recurrences are hypothesized to arise from disseminated tumor cells that reactivate after a long period of dormancy.

The survival, continued proliferation, and eventual formation of metastases by Disseminated tumor cells (DTCs) that have taken up residence in a secondary tissue may depend heavily on interactions with the surrounding microenvironment. A statistical analysis of the clinical manifestations of metastatic hazard supports a model of halted rather than continuous growth post-dissemination [11]. One possible explanation, supported by computational simulations, is that cells stop dividing for a period of time after lodging in a distal tissue (i.e., they become dormant) [12]. From a biological perspective, several mechanisms are proposed to be involved in the maintenance of dormancy [13, 14]. One particular example involves the differential activation of p38 MAPK and ERK1/2, downstream of various extra-cellular matrix (ECM) dependent interactions. Proliferation is generally considered to be increased by a high ERK1/2 to p38 activation ratio while the reverse may support a non-proliferative state [13, 14]. The activity of these growth regulating kinases is controlled in part by interactions between ECM proteins, extracellular regulatory factors such as plasmin, and transmembrane integrins which transmit these signals to the interior of the cell [13, 14]. These interactions that comprise the uPAR-integrin signaling network, which, in turn, can influence the balance of ERK1/2 and p38, affecting whether the cell remains quiescent or resumes proliferation [13, 14].

Studies suggest that breast cancer (BC) cells have certain attributes that make the bone marrow a particularly suitable niche for dissemination. Normal bone undergoes a continual process of renewal mediated by specialized cell types that remove (osteoclasts) and replace (osteoblasts) bone material [15]. This remodeling process is essential for the maintenance of healthy bone structure and calcium homeostasis (Clarke, 2008) [15]. Breast cancer cells that have metastasized to bone can interact with stromal cells in a manner that ultimately enhances metastatic growth [14, 16]. Typically, these interactions have an osteolytic effect leading to metastatic lesions and reduced bone density. As bony material is broken down factors such as TGFB, VEGF and FGF can be released into the surrounding environment, promoting proliferation and angiogenesis [14,16,17]. Evidence from mouse xenograft studies illustrates how BC cells can express osteomimetic factors that may contribute to their ability to persist in the bone microenvironment [17]. Twelve factors were identified in one meta-analysis that are known to be involved in functions such as proliferation and differentiation, cell adhesion, chemokine signaling, and bone resorption and remodeling [17]. In cell based assays, the presence of human bone marrow stem cells (hMSECs) or hMSC conditioned media drives the proliferation of ER+ tumor cell lines [18]. The Sasser study also showed that one of two ER- cell lines was stimulated albeit to lesser extent. In triple negative (TN) breast cancer, cancer-associated fibroblasts (CAFs) may select for primary tumor cells that respond to bone-derived chemokines and growth factors [19]. In TN breast cancer, expression of the 'stem-ness' associated genes ID1 and ID3 by primary tumor cells has been shown to be important for metastatic colonization and re-initiation in the lungs [20]. With these

examples in mind, it is evident that when and whether a disseminated tumor cell (DTC) resumes proliferation and forms a metastatic lesion is in many ways dependent on how it responds signals derived from the surrounding stroma. This leads to the question of whether, and to what extent the internal state of primary tumor cells influences their metastatic fate after dissemination. Evidence in the form of primary tumor biomarkers, and gene expression signatures seems to suggest that this is the case.

Over the last several decades, the standard of care for patients with ER+ tumors has been to administer anti-estrogen drugs such as Tamoxifen and aromatase inhibitors such as Letrozole [21, 22]. While these medicines can help prevent metastatic relapse, they are not without side effects, some of which may be quite serious including elevated risk of bone fracture due to osteoporosis, stroke, and endometrial cancer [21, 22]. Clinicians have a strong interest in being able to prioritize these therapies for patients who are most likely to benefit, without over-treating those who are unlikely to relapse [23]. Despite their prognostic value for early metastases, clinical parameters such as tumor size and grade are less helpful for late metastasis [23]. An alternative that has emerged is the search for signatures and biomarkers in expression profiles from primary tumors [23].

Breast Cancer Signatures

The emergence and maturation of standardized microarray technology has marked an important milestone in the progression of transcriptomic analysis. This, along with more recent advancements in high throughput sequencing of transcriptome

derived libraries (RNA-Seq), have made it possible to profile the activity of thousands of genes in a sample simultaneously. In the context of breast cancer, transcriptomic profiling has led to the identification of groups of genes or 'signatures' for which changes in the level at which signature genes are expressed corresponds strongly to a particular disease state or phenotype [23-28].

There are several motivations for breast cancer researchers to develop these signatures. These motivations include: gaining insight into the biological processes driving progression, improving approaches to estimating patient prognosis, and perhaps most importantly, to helping physicians make better informed treatment decisions [23]. Within the past two decades various groups have identified signatures that correspond to: tumor molecular subtype (including ER status), histological grade, and disease-free survival [23-28].

One approach to identifying a transcriptomic signature is to first assemble as large of patient cohort as possible that meets certain clinical and demographic criteria such as age, estrogen receptor status, or lymph node metastases [24-27]. Primary tumor tissue, harvested during therapy is processed and expression profiles are generated that represent an 'average' of mRNA transcripts present in the sample [24-28]. Data set size can range from tens to hundreds, and in some cases thousands of patients [24-29]. Having many patients may help to detect changes in gene expression that are related to the disease, as opposed to those that are due to factors such as age, diet, and environment.

Various statistical methods have been used to identify genes where observed changes are strongly associated with some aspect of the disease. These have included pairwise statistical tests [25], correlation with clinical markers [28], and various combinations of univariate and multivariate survival analyses [24, 27, 29]. For example; the Genomic Grade Index (GGI) was developed based on per-gene pairwise comparisons between profiles from Elston-Ellis grade 1 and grade 3 tumors [25]. The Sensitivity to Estrogen index was developed by selecting genes with expression profiles that correlated with expression of the estrogen receptor gene ESR1 [28].

A typical benchmark for signatures is the extent to which the signature can stratify patients into risk categories or otherwise divide patients into biologically meaningful groups. For example, the GGI was shown to distinguish between breast cancer molecular subtypes especially Luminal A and Luminal B [27]. This signature was also shown to distinguish between patients at higher risk of early metastasis [23, 25, 26]. An interesting observation with the 76-gene signature developed by Wang et al. was that patients with a low risk-score had a uniformly low rate of metastasis over 10 years whether or not they received adjuvant tamoxifen therapy [24, 30]. Patients with a high risk-score, on the other hand showed significant benefit from Tamoxifen in the first five years post intervention [24, 30].

One thing that several of the signatures developed between 2005 and 2009 have in common is that they are dominated by genes involved in the cell cycle, or somehow related to proliferation [23 – 30]. When Wang et el. developed their 76-gene signature specifically benchmarked their analyses against a 5-year window for

metastasis [24]. The above results seem to indicate collectively, that while patients whose primary tumors express high levels of certain proliferative genes are at greater risk for early metastatic recurrence, patterns in the expression of these genes is of limited predictive value for late metastatic events [23].

In their 2015 review article, Sestak and Cuzik discuss several gene expression based signatures considered to have prognostic value specifically for late distal metastases including Prosigna's PAM50 ROR, the Breast Cancer Index (BCI) developed by bioTheranostics, and EndoPredict (EP) and EndoPredict Clinical (EP-Clin) by Myriad Diagnostics (formerly Sividon) [23]. Of the 50 genes included in the PAM50 ROR [31], the 47 recognized by DAVID are enriched with terms related to proliferation such as "*cell-cycle*" and "*mitosis*". In addition, several genes in the PAM50 classifier are annotated with terms describing processes that occur in the bone microenvironment such as "*osteoblast differentiation"* and "*ossification".* The BCI includes 5 genes described by authors as being associated with the cell cycle, and 2 associated with Estrogen Receptor sensitivity [32]. The EndoPredict classifier consists of 8 genes [33], two of which also are included in the PAM50 ROR. While 8 is too few for enrichment analysis, an ontology search in DAVID shows these genes, are annotated with proliferation related terms such as "*cell cycle checkpoint"*, "*G2/M transition of mitotic cell*", and "*positive regulation of exit from mitosis*" and also terms related to the bone microenvironment such as "*ossification*", "*positive regulation of osteoblast differentiation*" and "*cellular calcium homeostasis*". Each of these signatures has been shown to add prognostic value to clinical information such as tumor size, hormone receptor status and lymph node status in estimating the likelihood

of distal recurrence greater than 5 years after initial diagnosis and treatment [23]. An implication of these findings is that genes involved in microenvironmental processes, and hormone sensitivity may be more informative for predicting late metastatic outcomes.

Several studies have pursued lines of transcriptomic inquiry with metastatic dormancy as a specific research goal. In one example, Kim et al., used genes associated with p38 dependent quiescence, or inhibited angiogenesis to define a 49-gene, directionally-specific signature consisting of 22 genes where upregulation and 27 genes where downregulation is attributed to a dormant state [34]. When these lists are analyzed in the current version of DAVID Functional Annotation Tool [35, 36], the upregulated gene list is enriched for terms relating to ECM interactions such as: "*cell migration*" and "*extracellular matrix organization*". The downregulated gene list from Kim et al. [34], is enriched for terms such as "*cell proliferation*" and "*regulation of cell cycle*". Authors benchmarked this signature in survival analyses using several publicly available datasets and were able to stratify ER+ patients in survival analyses.

Qing Cheng and colleagues assembled a "cohort" of 4767 patients using microarray data deposited in the public domain originally generated during the course of studies such as those undertaken by Loi et al., Sotiriou et al., and Wang et al. [24 – 29]. In a 743 patient subset, they identified gene expression clusters that stratified patients into groups having a good prognosis, early, or late metastatic disease [29]. Gene Set Enrichment analyses of patient profiles from the "late metastasis" sub group

G4 included ontology terms such as: "*regulation of cell differentiation*" and "*cell migration*". Along with changes in specific genes, these observations indicated the possibility that a reversal of the epithelial to mesenchymal transition may play a role in the escape from dormancy [29]. They further identified a 51-gene signature in tumor epithelium, enriched for the processes "*ECM remodeling*", "*fibrosis*" and "*EMT Transition*" [29].

With the objective of developing a signature specific for late metastatic events, Mittempergher and co-workers conducted a retrospective expression profiling analysis involving 252 frozen ER+ tumor specimens for which detailed clinical data was available [37]. Of particular interest is their division of patients into three groups based on whether or not they experienced metastatic disease within 10 years of diagnosis (M0: no relapse for at least 10 years follow up) and if so when (M5 <= 5 years < M5-15) [37]. They applied several algorithms to search for genes with prognostic profiles [37]. Using a supervised clustering approach, they extracted a 241-gene signature from a comparison between the M0 and M5-15 data set [37]. In 10-fold cross-validations, this signature had a 77% classification accuracy in predicting patients likely to experience a late metastatic event [37]. In analyses conducted using DAVID, the 241-gene classifier was enriched with terms such as "*extracellular matrix*" and "*immune response*" [37].

Introduction to the Key Gene Sets in Metastatic Dormancy project

The purpose of this study was to identify genes and pathways that may play a role in either entry into a dormant state, maintenance of such a state, or eventual

reactivation. In this project, the working hypothesis is that the genes expressed in primary tumor cells (PTCs) influence the fate of these cells once they have disseminated and may determine the likelihood of their reactivation. The central focus of this work was to identify patterns of gene expression that differentiated patients who experience late metastases from those who appear to avoid recurrence entirely.

The goal of this work was to extract a transcriptomic signature associated with persistent metastatic dormancy using publicly available primary tumor expression profiling data derived from female breast cancer patients.

The approach taken was to 1) find as many qualifying studies as possible, 2) obtain a list of differentially expressed genes for each of them, 3) analyze gene lists for enriched ontology terms, and 4) evaluate the extent to which analysis results from each study overlapped with one another. Consensus between several studies revealed biological processes that may be involved in the process of stromal-driven reactivation.

Chapter 2

STUDY AND PATIENT SELECTION

Overview and Selection Criteria

All of the data used in this work was retrieved from NCBI's Gene Expression Omnibus (GEO) between June of 2017 and January of 2018. The NCBI's Gene Expression Omnibus (GEO) is a database resource that archives microarray and high-throughput sequencing data [38]. While GEO accepts profiling data generated on a variety of platforms, the most common data type is gene expression profiling via microarray[38]. In 2012, GEO hosted just under 10000 series where gene expression was measured by microarray, and just under 1000 series where gene expression was measured by next generation sequencing [38]. The gap has narrowed substantially since then with 55,715 gene expression microarray series and 22,113 gene expression next generation sequencing series currently available (Appendix A). The following criteria were used to decide whether a dataset was suitable for inclusion:

- Maximum follow-up interval > 15 years
- Last metastatic event observed at least 12 years after harvest of primary tumor tissue.
- Samples for at least three ER positive patients in each analysis group.

Follow-up interval and time to metastases criteria were implemented with the intention of comparing tumor expression profiles from patients experiencing late distal metastases to profiles from patients where potentially indefinite dormancy is observed. Analysis was restricted to ER+ patients based on observations from previous

dormancy work [34, 37]. A minimum of three samples in each analysis group was necessary for pairwise statistics.

Four studies meeting the above criteria were identified for analysis. Each of the studies evaluated here is referred to by the last name of the primary author. Each study contributed 14 or more patients to the analysis, with at least 4 patients in each category (Table 1). For the Loi study, patients whose samples were analyzed on different chips (HG-U133A (Loi A), HG-U133 Plus 2 (Loi P)) were treated separately; however, the same cutoff was applied.

Dataset	GEO Accession	Late Patients	Never Patients	Cutoff (Years)	Maximum Follow Up (Years)
Loi A	GSE6532	15	4	13.4	16.9
Loi P	GSE6532	14	23	13.4	16.9
Zhang	GSE12093	7	7	13.5	15.8
Symmans	GSE17705	25	7	14.4	16.3

Table 1 Studies Selected for Evaluation

Microarray data from the Loi study was divided into two processing groups depending on the analysis platform (Loi A: HG-U133 A, Loi P: HG-U133 Plus 2)

Loi (Loi A and Loi P)

The data sets referred to here as "Loi A" and "Loi P" come from a set of patients whose breast tumors were treated at hospitals in Sweden and the United Kingdom [26]. Tumor samples collected between 1980 and 1995 were profiled on either the HG-U133A/B chip set or the combined HG-U133 Plus2 platform. [26, 27]. Patients from the Loi data set whose samples were analyzed on different chips (HG-U133A, HG-U133 Plus 2) were treated separately during differential expression and

enrichment analyses. In this work, the "Loi A" data set is the subset of patients analyzed on the HG-U133A platform, and the "Loi P" data set is the subset of patients analyzed on the HG-U133 Plus2 platform. All the Loi patients analyzed in this work had ER positive tumors (data accessible at NCBI GEO database [26, 27] accession GSE6532).

For the subset of patients from the Loi data set analyzed on the HG-U133A platform (Loi A), a total of 19 patients were identified that met analysis criteria. There were 15 patients that experienced a metastatic recurrence after 5 years and up to 12.5 years, and these patients were assigned to the "Late" group. In the "Late" group for Loi A, there were 9 patients that had received adjuvant tamoxifen therapy, and 6 patients that had not. There were four patients who were followed from 13.4 up to 16.9 years with no documented evidence of distal metastasis. These patients were assigned to the "Never" group. None of the "Never" patients received adjuvant tamoxifen therapy.

For the subset of patients from the Loi data set analyzed on the "HG-U133 Plus 2" platform (Loi P), there were 14 patients that experienced a distal metastatic event between 5 and 13.4 years and, accordingly, were assigned to the "Late" group. There were 23 patients with no observed metastases from 13.4 up to 16.9 years, and these patients were assigned to the "Never" group. All of the "Loi P" patients received adjuvant Tamoxifen therapy.

Zhang

The data set referred to here as "Zhang" comes from a set of 136 patients treated at several European institutions and one institution in the United States.

Microarray data was generated using frozen tissue specimens that had been collected between 1990 and 2000 [30]. All of the patients in this dataset had ER positive tumors and received adjuvant tamoxifen therapy. In this data set, the longest observed interval between initial therapy and metastatic recurrence was 13.5 years. There were 7 patients that experienced a distal metastatic event between 5 and 13.5 years and were assigned to the "Late" group. There were 7 patients with no observed metastases from 13.5 to 15.8 years follow-up ("Never" group). All of the patients in this data set were profiled on the HG-U133A platform (data accessible at NCBI GEO database [30], accession GSE12093).

Symmans

The data set referred to here as "Symmans" comes from a set of 298 patients treated in Austria, France and the United Kingdom between 1978 and 2002. In the original paper these patients are described as the "second validation cohort" [28]. All patients in this data set had ER positive cancer and received five years of adjuvant Tamoxifen therapy. In this data set, the longest interval between treatment of the primary tumor and metastatic was 14.4 years. There were 25 patients that experienced a distal metastatic event between 5 and 14.4 years follow-up. There were 7 patients with no observed metastases from 14.4 years up to 16.3 years follow-up. All of the patients in this data set were profiled on the HG-U133A platform (data accessible at NCBI GEO database [28], accession GSE17705).

Chapter 3

ANALYTICAL METHODOLOGY

Data Partitioning and Environment

Subjects were grouped based on whether they experienced a distant metastatic event within a defined time frame post diagnosis. For each study, the last recurrence (primary or metastatic) was used to define a cutoff dividing patients into two categories, "Late" or "Never". For each study, patients were assigned to the "Late" category if they experienced a metastatic event from 5 years after diagnosis until the last observed recurrence on that study. Patients with no observed recurrence after this cutoff were assigned to the "Never" group. Only data from patients with ER positive tumors was used for this analysis. In all of the studies evaluated, the expression profiles were from a sample of the patient's original primary tumor. All analyses were performed using R (3.2.3) for general purpose data processing, with Bioconductor (3.4) packages designed for specific bioinformatic analyses. All R scripts in the pipeline used to implement these analyses have been deposited in the following GitHub repository: https://github.com/afaranda/NeverLatePipeline (Appendix B).

Microarray Data Processing

Microarray technology is an established method for profiling gene expression in a wide variety of contexts and disciplines. In arrays used to measure gene expression, the array itself consists of a solid substrate onto which oligonucleotide probes are attached. Each probe is designed to correspond to a specific, known mRNA transcript. Messenger (m)RNA extracted from a biological specimen is amplified to prepare a cRNA library that can be analyzed on the array. The process of library preparation via in vitro transcription, often referred to as the Eberwine method,

was originally developed to overcome fidelity issues that affect PCR based amplification techniques [39]. A liquid sample containing a tagged cRNA library [40] is applied to the surface of the chip; library fragments hybridize with their corresponding probes; and the chip is interrogated with a laser scanner that excites the tagged fragments, which then emit a fluorescent signal. The more abundant a particular transcript, the "brighter" the corresponding signal. Probes are identified by their position in the array. In the Affymetrix HG-U133 series of microarrays, 25 nucleotide probes are used to measure the abundance of mRNA transcripts that are present in the library. The probes used on this series of microarrays were designed based on known mRNA transcripts recognized by build 133 of the UniGene transcriptomic database [41, 42]. There are several microarray platforms or "chips" that were developed by Affymetrix as part of this series. The HG-U133 set consists of 2 microarray chips designated HG-U133A and HG-U133B. The probes on the "A" chip are primarily designed to target fully sequenced mRNA transcripts, where the sequence includes the 3'UTR. The probes on the 'B' chip were designed predominantly based on 'expressed sequence tags' (ESTs), which are less robust. The HG-U133 Plus 2.0 chip was developed shortly after this set [41, 42]. Along with some additional probes, the "Plus 2.0" chip combines the "A" and the "B" probes on a single platform [41, 42]. In this work, for patients that were profiled using the "Plus 2.0" chip, only probes designed for the "A" chip were examined for differential expression and included in downstream enrichment analyses to enable direct comparison to data collected with "A" chips.

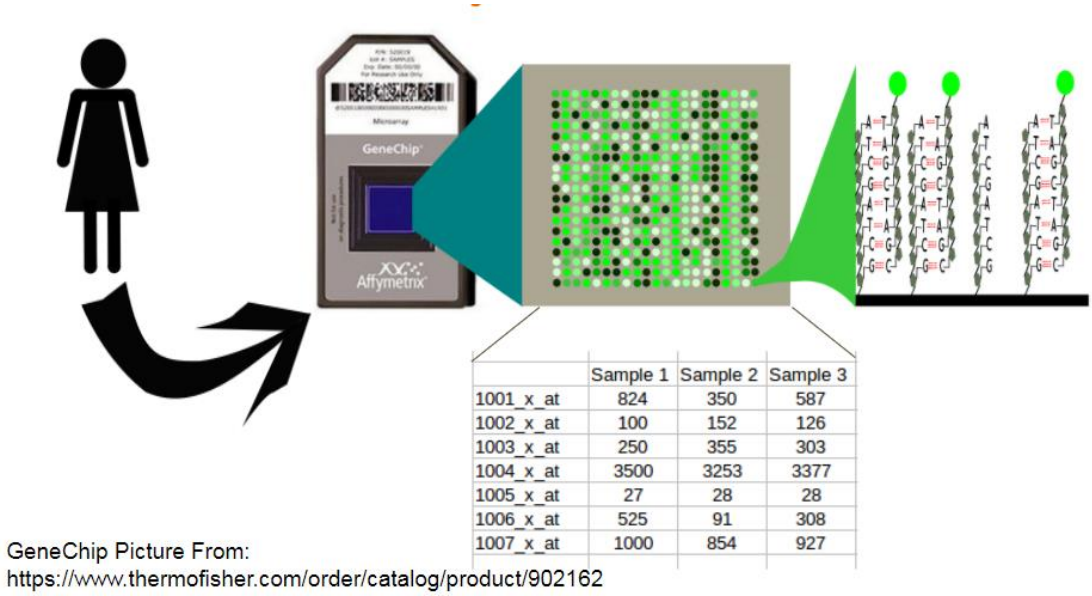

Figure 1 Microarray Analysis Workflow
In a typical workflow, mRNA isolated from patient's primary tumor would be processed into fluorescently labeled libraries and hybridized to arrays. Spot intensities are recorded by the array scanner.

Raw expression data, in the form of Affymetrix 'CEL' files, was retrieved from the Gene Expression Omnibus (GEO; Accession numbers GSE6532, GSE12093 and GSE17705). Each Cel file contains measured probe intensities for 22283 to 54675 probes. For each study, the Cel files from patients assigned to specific categories were assembled into an "affyBatch" for normalization and subsequent analysis of differential expression. Each batch (one batch per study) was processed individually. The matrix of probe intensities for each batch was normalized using the RMA algorithm [43]. For patients measured on the HG-U133 Plus 2 platform, only probes that were also detected by the HG-U133A platform were examined for differential expression or included in downstream enrichment analyses. Gene symbols

were assigned to Affymetrix probes based on annotation found in DAVID, as described below.

Quality Evaluation

After normalization, the matrix of probe intensities was evaluated for overall quality control. The package 'arrayQualityMetrics' was used to prepare figures and calculate statistics [44]. Patients that appeared to be outliers were flagged. The implementation used provided outlier detection based primarily on the following three metrics: the L1 Distance between arrays, goodness-of-fit against a pooled distribution of probe intensities, and Hoeffding's D to evaluate the quality of individual arrays [44].

Differential Expression

The Limma algorithm was used to identify genes that may be differentially expressed between the two patient groups [45]. For each study, probes were selected for enrichment analysis if the magnitude of the fold change between "Never" and "Late" patients was greater than 1.5 and had an associated *p*-value less than 0.05. Overlap between lists of differentially regulated genes was evaluated using R's built-in set operation and tabulation functions.

Enrichment Analysis

Since DAVID is capable of processing Affymetrix probe ID's, lists of differentially expressed probes were submitted directly to DAVID for annotation and enrichment analysis [35, 36]. In order to capture as much pathway perturbation as

possible, gene lists were not partitioned based on the direction of fold change. The lists submitted to DAVID included probes with positive and negative fold changes. The process of querying DAVID was automated using the bioconductor package 'RDAVIDWebservice' [46]. This package can be used to open a connection to the DAVID Knowledgebase and perform various analytical tasks such as gene list submission and retrieval of enrichment results.

For terms from the Gene Ontology (GO), DAVID provides eight annotation categories that the user can select during enrichment analyses. Five of these were designed based on the depth of term in the GO hierarchy, and are assigned a numerical designation (Levels 1 - 5) [35, 36]. There are three additional categories designated "ALL", "DIRECT" and "FAT". As the name implies, the Category "ALL" includes all terms in a particular ontology branch (Biological Process, Molecular Function, or Cellular Component) (Appendix C, Personal communication). The "DIRECT" category restricts term mappings to those assigned directly to a gene by an annotating resource such as Uniprot (Appendix C, Personal communication). The category "FAT" is a subset of the Gene Ontology consisting primarily of higher level terms that have been filtered to remove broadly defined terms and focus on more specific ones. In the DAVID knowledgebase, the category "GOTERM_BP_FAT" is the FAT categorization of the biological process branch. While data was collected for several categories, this was the primary category that was considered for comparisons between studies and used for biological evaluation. Terms from the category "GOTERM_BP_FAT" were considered enriched if their Benjamini-Hochberg

adjusted EASE score was less than 0.05. Overlap between lists of ontology terms was evaluated using R's built-in set operation and tabulation functions.

The enrichment summarization tool REVIGO was used to obtain a high-level view of process enrichments from each data set individually and for overlapping terms [47]. Analysis was completed manually by uploading lists of Terms with their corresponding Benjamini-adjusted EASE scores to the REVIGO server accessed via http://revigo.irb.hr/. The default settings were used for all analyses (Table 2).

Table 2 Default parameters for REVIGO analyses

Parameter	Setting
Allowed similarity	Medium (0.7)
Numbers associated with GO Terms	"p Values"
Select a database with GO term sizes	Whole UniProt (default)
Select a semantic similarity measure to use	SimRel

Randomized Probe Lists

A set of 1000 randomly generated probe-lists was generated by sampling 500 probe id's at a time, without replacement, from the 22283 unique ID's for probes that are present on the HG-U133A microarray. Each list was submitted to DAVID via an RDAVIDWebservice Query. The DAVID server limits the number of queries that can originate from any one source to 200 per day. In order to obtain results in a timely manner, the set of probe-lists was divided into small batches that could be submitted in parallel from multiple devices. For each of 1000 lists, the set of enriched terms was was retrieved for the category GOTERM_BP_FAT. Once tabulated, these data were used to calculate summary statistics and prepare histograms.

Chapter 4

DIFFERENTIAL GENE EXPRESSION AND PROCESS ENRICHMENT

Quality analysis using the array quality metrics package

In the normalized data set for patients from the Loi study analyzed on the HG-U133A platform (Loi A), one sample from the 'Late' group was flagged as an outlier based on its L1 distance from the remaining arrays. In addition to probes that detect mRNA transcripts, microarrays produced by Affymetrix also include a number of quality control probes. None of these microarray quality control probes were differentially expressed between the two comparison groups; therefore, this sample was not considered to have significantly impacted the analysis. All of the samples in the original batch were retained (Appendix D).

In the normalized data set for patients from the Loi study analyzed on the HG-U133 Plus2 platform (Loi P), 3 samples from the 'Never' group were flagged as outliers by array quality metrics. One sample was flagged as an outlier based on its distribution of probe intensities and its L1 distance from the remaining samples. The two others were flagged based on distance alone. None of the microarray quality control probes were differentially expressed between the two comparison groups; therefore, these samples were not considered to have significantly impacted the analysis. All of the samples in the original batch were retained (Appendix D).

In the normalized data set for Zhang, one sample was flagged as an outlier based on its distribution of probe intensities and its Euclidean distance from the other samples. Two Affymetrix QC probes were detected as differentially expressed (Table

3). Both differentially expressed QC probes detect human 18s ribosomal RNA present in the tissue sample. When this outlier is removed, there are two other samples that become classified as outliers. After five successive rounds of outlier removal and normalization, it would have been necessary to drop 6 samples in order to have a Zhang data set that was free of outliers(Data not shown). Rather than lose nearly half of the samples, all samples meeting "Never / Late" criteria were retained (Appendix D).

In the normalized data set for Symmans, there were four samples that were flagged as outliers. Two of these were flagged based on their L1distance from the other samples and their probe intensity distributions. The other two were flagged based on their distance only. Samples from the Symmans study were analyzed at two different laboratories, labeled "JBI" and "MDA" in the data set. In this case, the four outliers were analyzed at JBI; all remaining samples were analyzed at MDA. It is well established that technical variation between laboratories can influence microarray results; therefore, the JBI samples were removed from consideration (Appendix D). Even with the removal of these outliers, there were 8 Affymetrix QC probes with differential measurements (Table 3).

Table 3 Affymetrix quality control probes with differential mean intensity between the two sample groups.

Target and Origin information was retrieved from the "NetAffx" database provided by Affymetrix [48].

Study	Probe	Target	Origin	Log_2 Fold Change	*p* Value
Zhang	AFFX-r2-Hs18SrRNA-3_s_at	Ribosomal RNA	Endogenous to sample	1.52	0.030
	AFFX-HUMRGE/M10098_3_at	Ribosomal RNA	Endogenous to sample	1.36	0.032
Symmans	AFFX-HSAC07/X00351_5_at	Actin, beta	Endogenous to sample	-1.04	0.003
	AFFX-HSAC07/X00351_M_at	Actin, beta	Endogenous to sample	-0.75	0.009
	AFFX-r2-Ec-bioC-3_at	*E. coli* biotin synthase	Spike in Control	0.78	0.021
	AFFX-BioC-3_at	*E. coli* biotin synthase	Spike in Control	0.79	0.025
	AFFX-r2-Ec-bioC-5_at	*E. coli* biotin synthase	Spike in Control	0.79	0.028
	AFFX-BioC-5_at	*E. coli* biotin synthase	Spike in Control	0.70	0.030
	AFFX-r2-Hs18SrRNA-M_x_at	Ribosomal RNA	Endogenous to sample	-0.76	0.036
	AFFX-HUMRGE/M10098_3_at	Ribosomal RNA	Endogenous to sample	-1.40	0.039

Differential Expression

For the purposes of enrichment analysis, genes were considered differentially expressed if the average fold change was greater than 1.5, with a Limma *p*-value less than 0.05. In "Never vs. Late" pairwise comparisons, samples from "Late" group patients were considered to represent the baseline condition. In "Never" samples, genes were considered upregulated when expressed at higher levels, and down regulated when expressed at lower levels.

Table 4 Summary of Differentially Expressed Genes by Study

(Minimum adjusted *p*-value: Benjamini-Hochberg adjustment of Limma *p*-values.)

Study	Affymetrix Probes	DAVID Genes	Upregulated	Downregulated	Min. *p* Value	Min. Adjusted *p* Value
Loi A	664	564	213	351	1.86E-05	0.22
Loi P	518	415	318	97	1.29E-05	0.17
Zhang	511	411	343	68	1.28E-05	0.23
Symmans	136	106	46	60	3.96E-05	0.55

For samples from patients in the Loi study profiled on the HG-U133A chip (Loi A), 664 probes were differentially expressed, corresponding to 564 genes recognized by DAVID. There were 213 genes upregulated in the "Never vs. Late" comparison, and 351 genes downregulated in the "Never vs. Late" comparison (Table 4).

For samples from patients in the Loi Study profiled on the HG-U133 Plus2 chip, 518 probes were differentially expressed, corresponding to 415 DAVID genes. There were 318 genes upregulated in the "Never vs. Late" comparison samples, and 97 genes downregulated in the "Never vs. Late" comparison (Table 4).

In the Zhang dataset, 511 probes corresponding to 411 DAVID genes were differentially expressed. There 343 genes upregulated in "Never vs. Late" comparison and 68 genes downregulated in the "Never vs. Late" comparison (Table 4).

In the Symmans dataset, 136 probes were differentially expressed corresponding to 106 DAVID genes. There were 46 genes "Never" samples and 60 genes expressed at higher levels on average in "Late" samples (Table 4).

Overlaps in Differentially expressed genes

The Jaccard index can be used to evaluate the relative similarity between a pair of lists [49]. It is calculated by enumerating the intersection of the two lists and the union of the two lists then dividing the number of elements in common by the total number of unique elements. A Jaccard index of 1 indicates that the two lists are identical. A Jaccard index of 0 indicates that there are no elements (genes) in common. For all pairwise comparisons between studies, the Jaccard indices of gene lists fell between 0.02 and 0.08 (Table 5). For example, there were 72 genes that were differentially expressed in both the Loi A patients and the Zhang patients. The union of gene lists for these two studies consists of 903 genes; therefore, the Jaccard index for these two studies is 72 divided by 903, or approximately 0.08.

It was noteworthy that for genes that were differentially expressed in more than one data set, there were many cases where the direction of change was contradictory between any two data sets. For example, consider changes in the expression of JUN between "Late" and "Never" patients, as they were observed in Loi A and Zhang. In the Loi A patients, JUN is expressed on average at a higher level by

"Never" patients than by "Late" patients. In Zhang patients, the reverse is true; JUN is expressed on average at a higher level by Late patients.

We evaluated the scope of these inconsistencies by tabulating cases of agreement, where in both studies the gene's expression changes in the same direction, and disagreement, where the direction of change in a given gene is opposite between two studies, for all study pairs. For example, when comparing Loi A patients to Loi P patients, we find that there are 65 genes that are differentially expressed in both data sets (Figure 2); however, there were only 11 in agreement, and there were 54 genes that changed in opposite directions for each study (Table 5). In comparing Loi A to Symmans, there were 10 genes in agreement and four which disagreed (Table 5). In comparing Loi A to Zhang, there were 10 genes in agreement and 62 which disagreed. In comparing Loi P to Symmans, there were two genes in agreement and 6 that disagreed. The greatest agreement was observed between Loi P and Zhang, where 23 of 32 overlapping genes were in agreement. The least was between Loi A and Zhang, where only 10 of 72 genes were in agreement (Table 5). Of the four data sets evaluated Loi P and Zhang showed the greatest similarity to one another with respect to differential expression; Loi A and Symmans were also more similar to one another than they were to the other two studies.

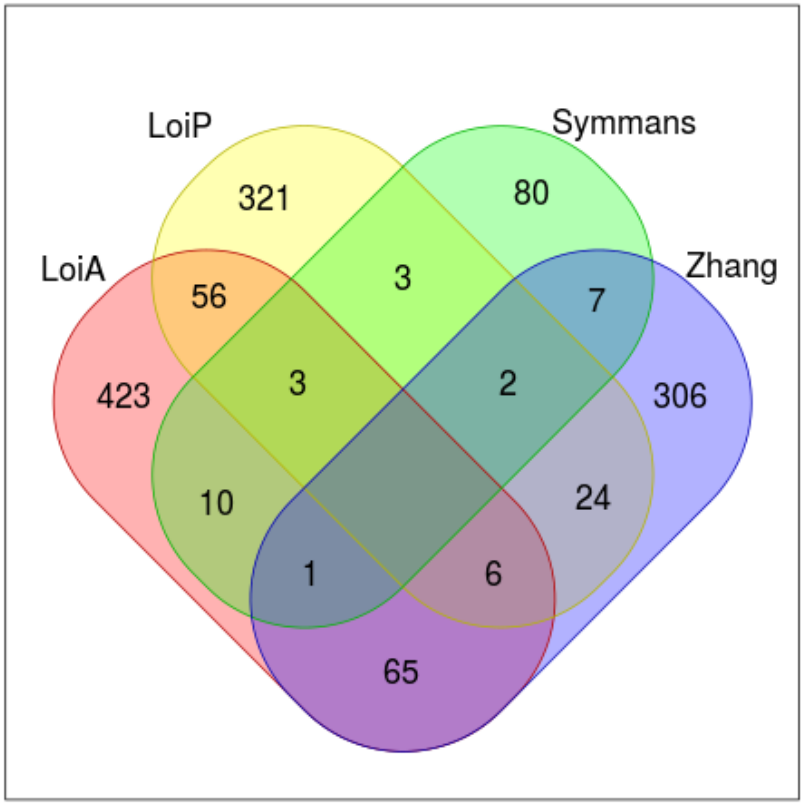

Figure 2 Venn diagram illustrating overlap in sets of differentially expressed genes from each of the four studies.

Table 5 Overlap and Directional Agreement in Differential Expression between four studies.

For each pair of studies, the following are tabulated: the number of genes in common (Total), the number genes that are differentially expressed in the same direction (Agree), and the number of genes that are differentially expressed in opposite directions (Disagree).

Study Pair	Total	Jaccard Index	Agree	Disagree
Loi A – Loi P	65	0.07	11	54
Loi A – Symmans	14	0.02	10	4
Loi A – Zhang	72	0.08	10	62
Loi P – Zhang	32	0.04	23	9
Loi P – Symmans	8	0.02	2	6
Symmans – Zhang	10	0.02	7	3

Ontology Enrichment

In the set of 564 differentially expressed genes from the "Loi A" samples, there were 484 ontology terms from the DAVID category "GOTERM_BP_FAT" that were enriched with a Benjamini adjusted EASE score < 0.05. When ranked by their Benjamini adjusted EASE score, the top three terms were 1) response to organic substance (9.52×10^{-17}), 2) extracellular structure organization (8.02×10^{-16}), and 3) extracellular matrix organization (1.06×10^{-15}). REVIGO Semantic similarity analysis reduced this set to 91 distinct terms divided into four major clusters (groups of more than 3 terms) (Figure 3).

- The cluster labeled "*response to organic substance*" consisted of 25 terms, including terms such as "*response to wounding*", "*STAT cascade"*, and "*response to oxidative stress*".
- The cluster labeled "*circulatory system development*" consisted of 20 terms, including terms such as "*collagen metabolic process*", "*ossification*", and "*tissue migration*".

- The cluster labeled "*regulation of cellular protein metabolism*" consisted of 19 terms, including terms such as "*protein phosphorylation*", "*regulation of cell proliferation*", and "*regulation of ERK1 and ERK2 cascade*".

- The cluster labeled "*extracellular matrix organization*" consisted of 17 terms, including terms such as "*extracellular structure organization*", "*cell migration*", and "*localization of cell*".

In the set of 415 differentially expressed genes from the "Loi P" samples, there were 676 ontology terms from the DAVID category "GOTERM_BP_FAT" that were enriched with a Benjamini adjusted EASE score < 0.05. When ranked by their Benjamini adjusted EASE score, the top three terms were 1) immune response ($6.76x10^{-57}$), 2) regulation of immune system process ($7.97x10^{-48}$), and 3) positive regulation of immune system process ($7.26x10^{-45}$). REVIGO Semantic similarity analysis reduced this set to 57 distinct terms divided into four major clusters (groups of more than 3 terms) (Figure 4).

- The cluster labeled *"immune response*" consisted of 21 terms, including terms such as "*inflammatory response*", "*cellular response to cytokine stimulus*", and "*regulation of type 2 immune response*".

- The cluster labeled "*cell activation*" consisted of 12 terms, including terms such as "*localization of cell*", "*cell migration*", and "*phagocytosis*".

- The cluster labeled "*protein phosphorylation*" consisted of 9 terms, including terms such as "*ERK1 and ERK2 cascade*", "*regulation of protein metabolic process*", and "*cytokine metabolic process*".

- The cluster labeled "*regulation of cytokine production*" consisted of 7 terms, including terms such as "*cytokine production*", "*regulation of angiogenesis"*, and "*regulation of leukocyte mediated cytotoxicity*".

In the set of 411 differentially expressed genes from the "Zhang" samples, there were 640 terms from the DAVID category "GOTERM_BP_FAT" that were enriched with a Benjamini adjusted EASE score < 0.05. When ranked by their Benjamini adjusted EASE score, the top three terms were 1) regulation of multicellular organismal development ($7.74x10^{-13}$), 2) cardiovascular system development ($8.26x10^{-13}$), and 3) circulatory system development ($8.26x10^{-13}$). REVIGO Semantic similarity analysis reduced this set to 107 distinct terms, divided into three major clusters (Figure 5).

- The cluster labeled "*regulation of cell proliferation*" consisted of 54 terms, including terms such as "*positive regulation of cell-cycle process*", "*MAPK cascade*", and "*inflammatory response*"
- The cluster labeled "*circulatory system development*" consisted of 31 terms, including terms such as "*cardiovascular system development*", "*ossification*", and "*cytokine production*".
- The cluster labeled "*cell motility*" consisted of 15 terms, including terms such as "*extracellular structure organization*", "*localization of cell*", and "*cell junction organization*".

In the set of 106 differentially expressed genes from the "Symmans" samples, there were 31 ontology terms from the DAVID category "GOTERM_BP_FAT" that were enriched with a Benjamini adjusted EASE score < 0.05. When ranked by their Benjamini adjusted EASE score, the top three terms were 1) response to organic substance ($3.49x10^{-03}$), 2) macromolecule localization ($6.70x10^{-03}$), and 3) cell death ($1.92x10^{-02}$). REVIGO Semantic similarity analysis reduced this set to 14 distinct terms, with one major cluster consisting of 7 terms. This cluster was labeled "negative regulation of protein metabolism" (Figure 6).

When intersections between term lists were evaluated, it was found that 180 terms from the DAVID category GOTERM_BP_FAT were enriched in 3 or more studies. A ranking statistic was used to order these terms for REVIGO analysis. For each term, the median of the Benjamini adjusted EASE scores from each dataset where the term was enriched was divided by the number of datasets where the term was enriched. For example, the term "programmed cell death" was enriched in expression data from all four studies. The median of the Benjamini adjusted EASE scores across studies (Loi A: 1.42×10^{-5}, Loi P: 1.00×10^{-5}, Symmans: 2.37×10^{-2}, Zhang:1.32×10^{-7}, MEDIAN: 3.03×10^{-6}) was divided by four giving a ranking score of 7.58×10^{-7}. REVIGO Semantic similarity analysis reduced this set to 51 distinct terms, with three major clusters (Figure 7).

- The cluster labeled "*Positive regulation of multicellular organismal process*" consisted of 23 terms, including "*regulation of immune system process*", "*regulation of apoptotic process*", and "*cytokine production*"
- The cluster labeled "*cellular response to organic substance*" consisted of 14 terms, including "*response to endogenous stimulus*", "*inflammatory response*", and "*wound healing*".
- The cluster labeled "*cell motility*" consisted of 8 terms, including terms such as "*localization of cell*", "*movement of cell or subcellular component*", and "*cell activation*".

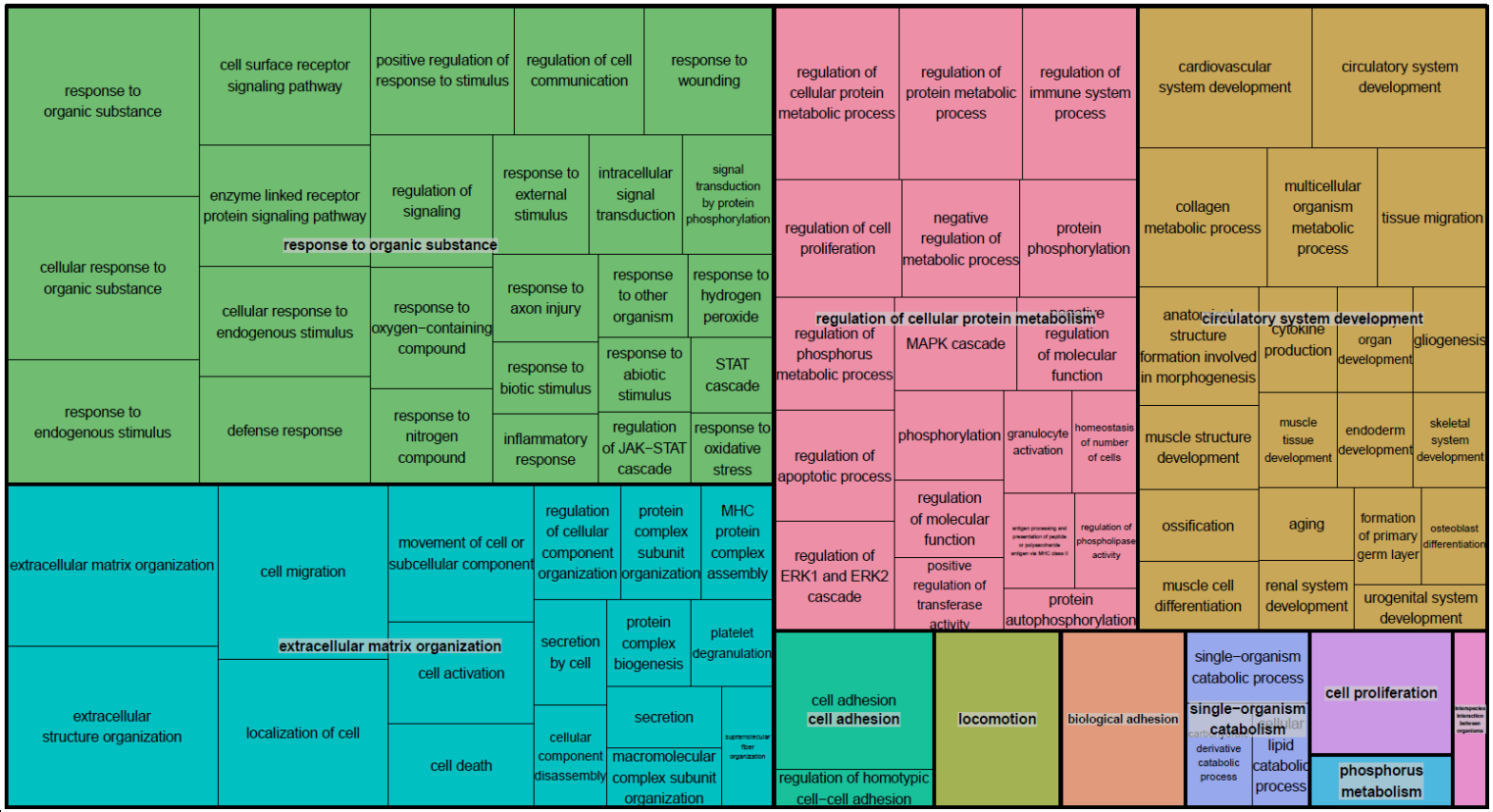

Figure 3 Loi A REVIGO Treemap

The size of each rectangle is determined by the enrichment score (Benjamini adjusted EASE score) of the representative term. Related terms are grouped into larger, color-coded clusters.

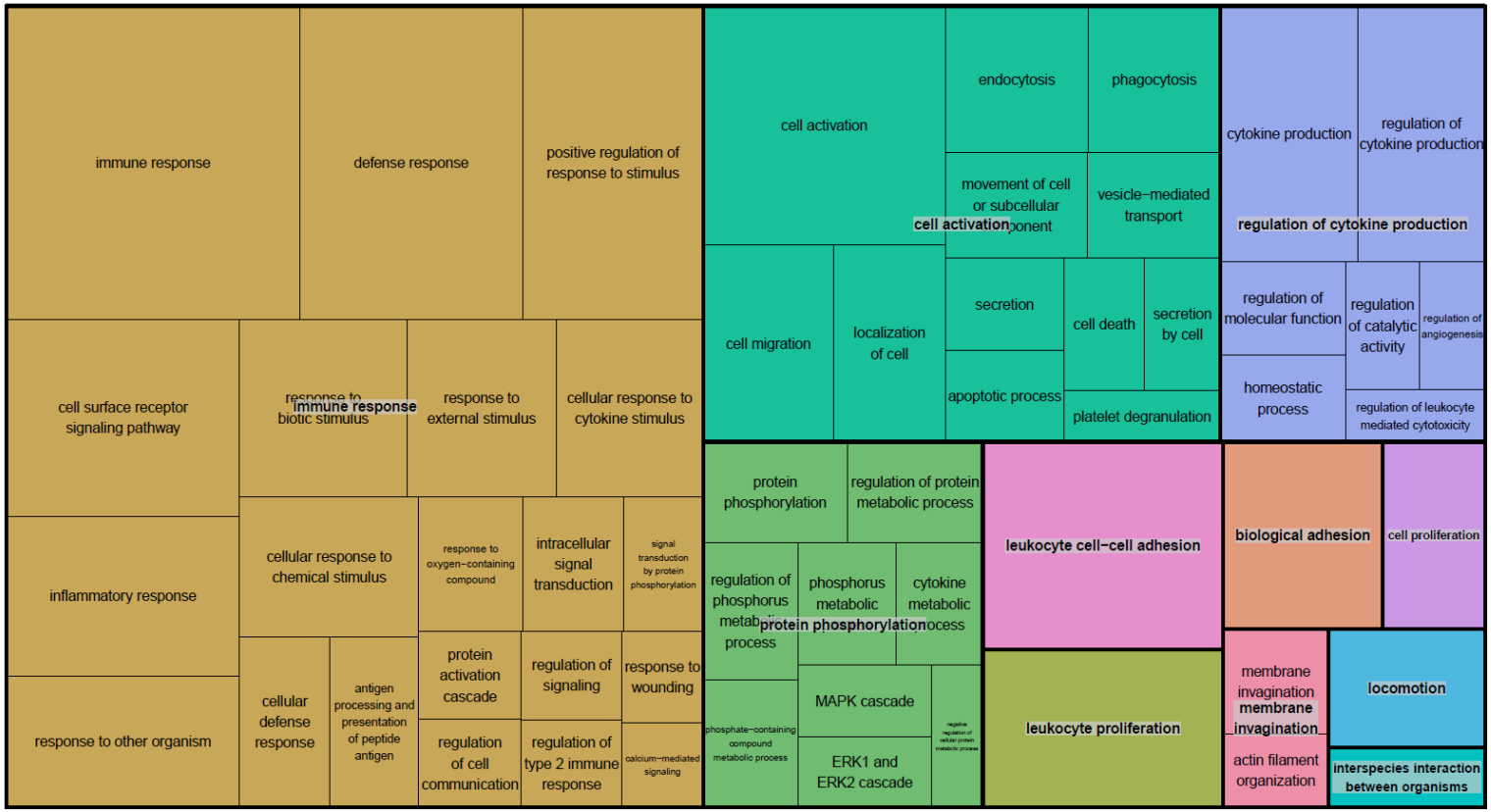

Figure 4 Loi P REVIGO Treemap

The size of each rectangle is determined by the enrichment score (Benjamini adjusted EASE score) of the representative term. Related terms are grouped into larger, color-coded clusters.

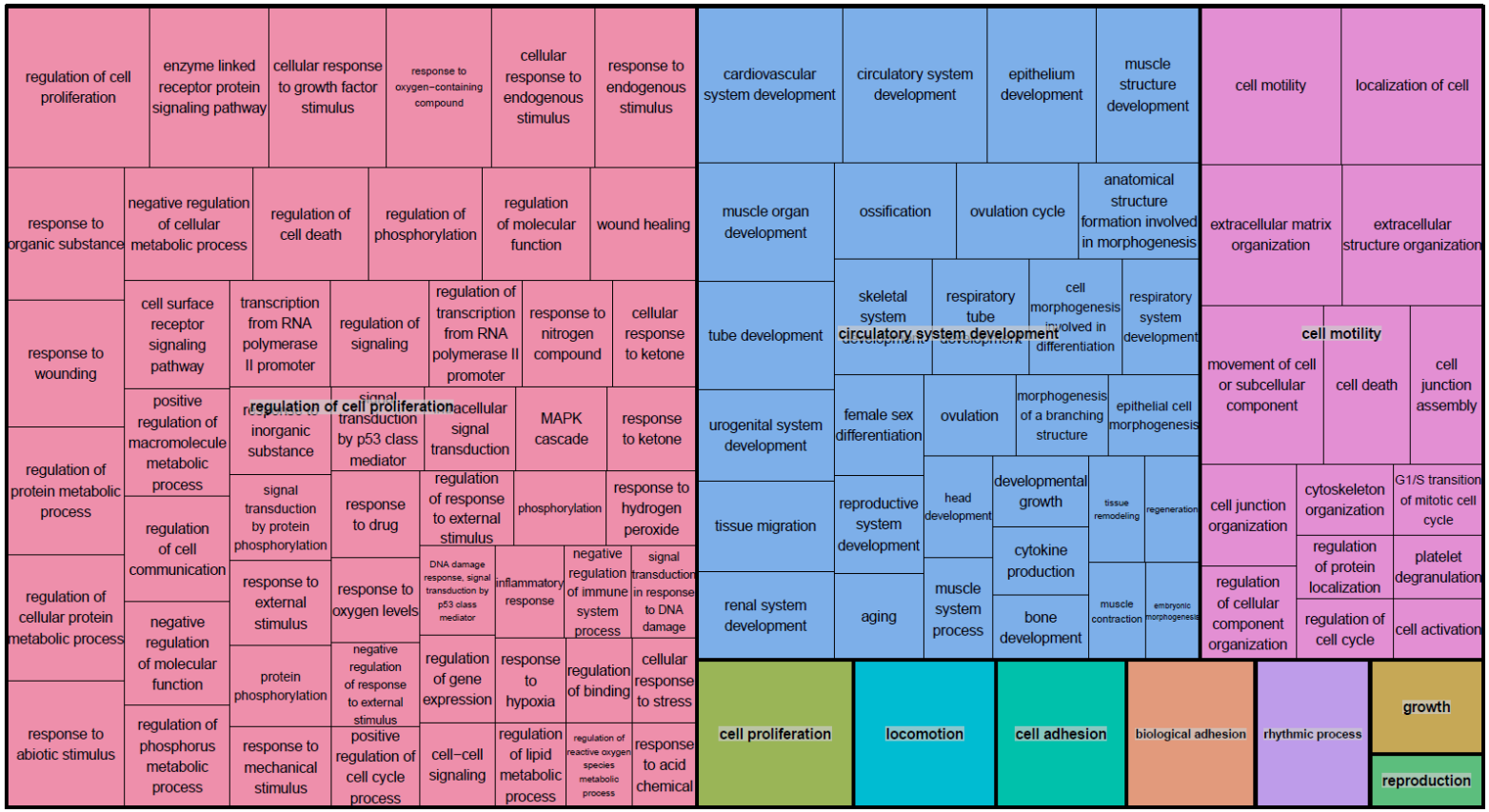

Figure 5 Zhang REVIGO Treemap

The size of each rectangle is determined by the enrichment score (Benjamini adjusted EASE score) of the representative term. Related terms are grouped into larger, color-coded clusters.

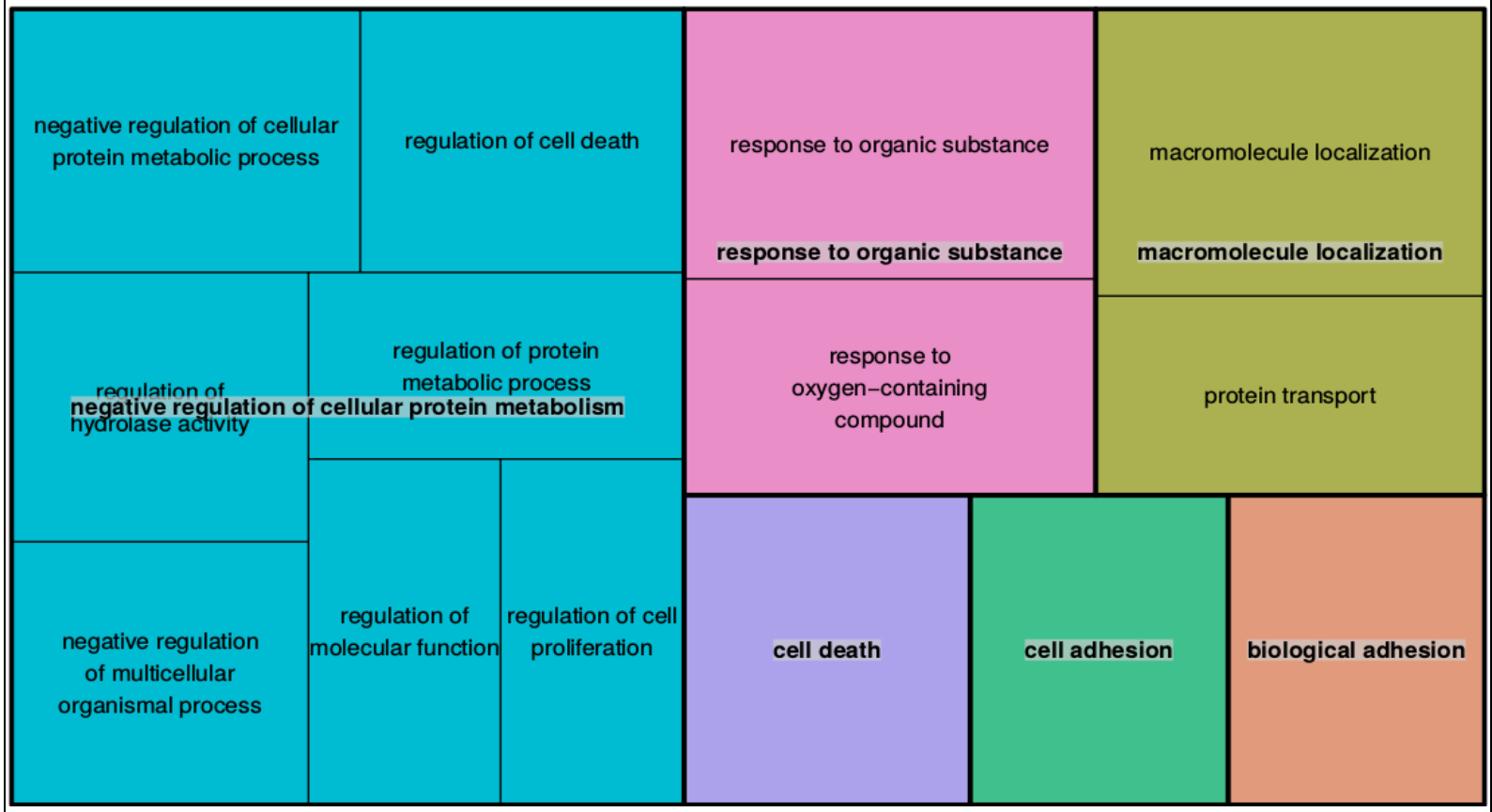

Figure 6 Symmans REVIGO Treemap
The size of each rectangle is determined by the enrichment score (Benjamini adjusted EASE score) of the representative term. Related terms are grouped into larger, color-coded clusters.

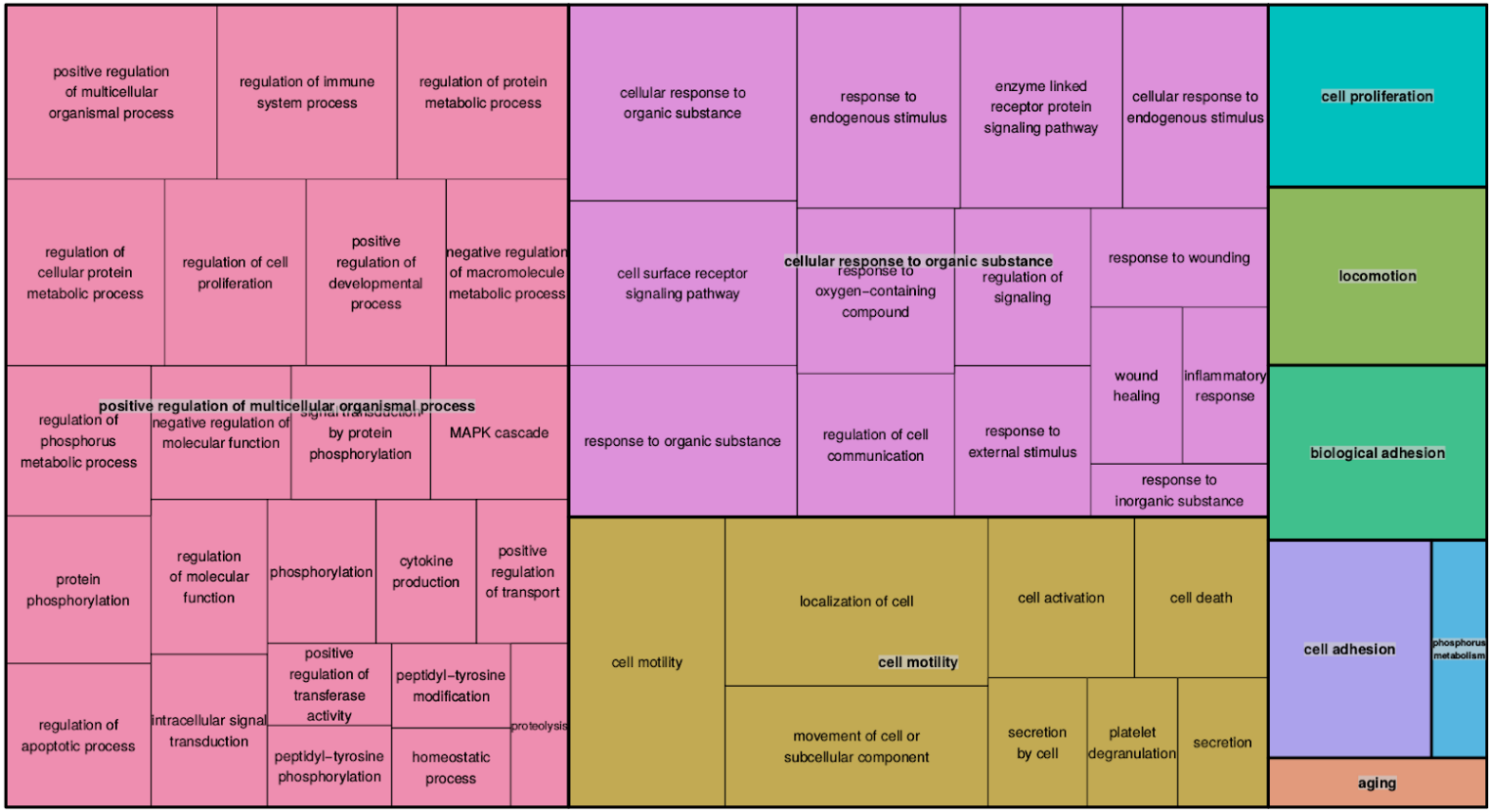

Figure 7 Overlapping Terms REVIGO Treemap
The size of each rectangle is determined by the enrichment score (Benjamini adjusted EASE score) of the representative term. Related terms are grouped into larger, color-coded clusters.

Evaluation of Random Gene lists

The results of pair-wise differential expression analyses showed only weak statistical significance. After adjustment for false discovery, the strongest p-value observed for any study was 0.17. In addition, the frequency of directional disagreement observed with genes differentially expressed in more than one study diminished, to a degree, the biological significance of these results. Nevertheless, enrichment results were remarkably strong. For all four studies, a large number of gene ontology terms were enriched with Benjamini-adjusted EASE scores well below 0.05.

This contrast between weaknesses in the differential expression and the relative strength of enrichment results raised the question of the likelihood of obtaining similar enrichment results from a list of random genes. The purpose of analyzing randomized probe lists was to develop a 'null' model against which the results from breast cancer patients could be compared toward assessing if such significant enrichment results could be observed from a list of random genes. Two major criteria were considered in this comparison: the best (minimum, Benjamini-adjusted) EASE score for each of the 1000 probe lists submitted to DAVID and the number terms enriched with an adjusted EASE score < 0.05 in each list.

To facilitate visualization, a negative log 10 transformation was applied to the best EASE scores in the category GOTERM_BP_FAT (Table 6, Figure 8). The distribution best EASE scores (minimum, Benjamini adjusted, and negative Log10 transformed) was a right-skewed bell curve with an average "best adjusted EASE score" of 0.083 and a median "best adjusted EASE score" of 0.023. The overall best adjusted EASE score was 1.30x10-7.

Table 6 Best adjusted EASE score with Negative Log_{10} transformation for each breast cancer dataset

Dataset	Min. Adj. EASE (Neg.Log_{10})
Loi A	$9.52x10^{-17}$ (16.0)
Loi P	$7.97x10^{-48}$ (47.1)
Zhang	$7.74x10^{-13}$ (12.1)
Symmans	$6.7x10^{-3}$ (2.2)

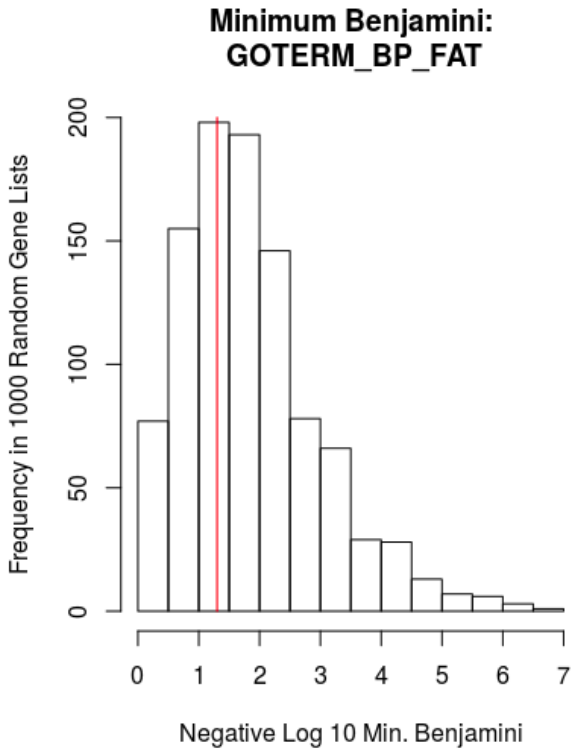

Figure 8 Distribution of minimum Benjamini adjusted EASE scores (after negative log 10 transformation) for terms in the category GOTERM_BP_FAT from 1000 random probe lists submitted to DAVID. The red line indicates the threshold of statistical significance.

For all 1000 of the lists of random probes submitted to DAVID, the DAVID server recognized at least 90% of the probe ids on the list. Between 14 and 47 probes were not mapped to DAVID genes. For each random probe list, between 0 and 217 terms from the category GOTERM_BP_FAT were significantly enriched. Out of the 1000 random probe lists (n = 500 probes), 11 were enriched with more than 150 terms, and 3 were enriched with more than 200 terms. For the entire set of lists, 18608

redundant terms from the category GOTERM_BP_FAT and were significantly enriched (Benjamini adjusted EASE score < 0.05).

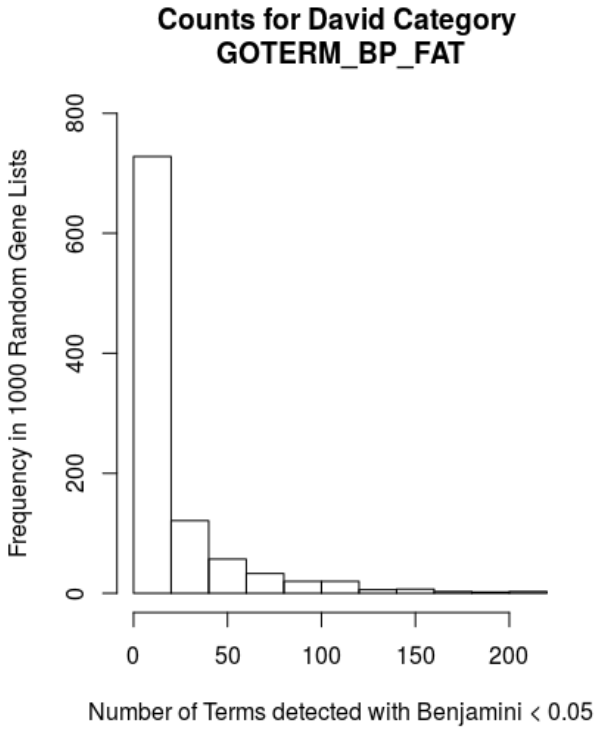

Figure 9 Distribution of the number of terms significantly enriched in random lists. Out of 1000 lists, 892 generated 50 or fewer hits.

Of the 18608 redundant 'hits' from the category GOTERM_BP_FAT there were 1614 unique ontology terms. For example, out of 1000 random lists, the term "*response to organic substance*" was enriched in 282 lists; the term "*programmed cell death*" was enriched in 125 lists; and the term "*ossification*" was enriched in 10 lists. Overall, there were 6 terms that were enriched in 200 or more random lists (> 20%) and 37 terms that were enriched in 100 or more random lists (> 10%). There were

1076 terms that were enriched in five or fewer lists. In the set of terms enriched in 3 or more studies, many of the top ranked terms were enriched in 20 or more random gene lists (Table 7).

Based on our random sample, the probability of obtaining 200 or more significantly enriched processes from the category GOTERM_BP_FAT is estimated to be 0.3%. Importantly and in contrast, for the breast cancer data sets, Loi A, Loi P and Zhang, over 400 terms from GOTERM_BP_FAT category were enriched. Likewise, for these three data sets, the minimum observed Benjamini adjusted EASE score was at least 7 orders of magnitude smaller than 1.30x10-7, the lowest observed by chance. The exception to these was the Symmans data set. In comparison to randomly generated enrichments Symmans ranked 239th based on the minimum observed Benjamini adjusted EASE score, and 32nd based on the number of significant terms.

Table 7 Top 30 Terms enriched in three or more studies.

Terms ranked by the scoring metric used for submission to REVIGO. The column labeled 'Hit' refers to the number of times the term was enriched in a random gene list (out of 1000). The Benjamini adjusted EASE score is reported for each dataset (Symmans abbreviated 'Sym.'). Values in the Rand* column are the minimum Benjamini adjusted EASE scores from random probe lists.

		Benjamini Adjusted EASE Score				
Term	Hit	Rand*	Loi A	Loi P	Sym.	Zhang
cell migration	54	8.24E-05	1.11E-13	1.20E-15	NA	6.09E-12
cell motility	50	5.75E-05	1.31E-11	2.56E-13	NA	8.67E-13
cellular response to organic substance	209	5.55E-06	1.10E-14	4.40E-13	NA	3.21E-09
localization of cell	50	5.75E-05	1.31E-11	2.56E-13	NA	8.67E-13
cell proliferation	82	4.71E-05	4.57E-08	1.42E-11	NA	9.29E-13
locomotion	59	1.34E-04	3.33E-10	4.16E-11	NA	6.55E-12
biological adhesion	88	1.40E-05	4.76E-10	3.32E-18	2.32E-02	6.61E-11
cell surface receptor signaling pathway	129	4.64E-05	5.87E-11	2.81E-28	NA	1.47E-06
positive regulation of multicellular organismal process	95	1.99E-04	4.79E-08	7.93E-11	NA	3.10E-10
regulation of cell motility	30	5.88E-05	1.38E-08	1.52E-11	NA	2.56E-10
regulation of cell migration	27	4.90E-05	9.41E-10	3.64E-12	NA	2.45E-10
regulation of locomotion	29	5.18E-05	2.19E-08	2.12E-11	NA	3.25E-10
cell adhesion	84	2.39E-05	3.91E-10	2.67E-18	2.40E-02	6.24E-11
regulation of cellular component movement	36	4.47E-05	3.61E-08	3.46E-10	NA	1.87E-10
response to organic substance	282	3.95E-07	9.52E-17	6.10E-13	3.49E-03	5.02E-09
response to endogenous stimulus	141	6.20E-06	5.07E-11	1.09E-02	NA	1.62E-09
enzyme linked receptor protein signaling pathway	82	3.04E-04	1.34E-09	1.17E-02	NA	6.27E-11
regulation of multicellular organismal development	84	1.11E-04	1.53E-09	4.12E-05	NA	7.74E-13
cellular response to chemical stimulus	250	1.10E-06	5.66E-15	3.30E-15	3.84E-02	1.95E-08
movement of cell or subcellular component	68	1.54E-05	3.64E-09	6.45E-09	NA	3.12E-11
regulation of immune system process	30	1.16E-03	4.24E-09	7.97E-48	NA	2.04E-02
regulation of protein metabolic process	128	1.41E-05	8.49E-10	4.40E-08	3.76E-02	1.06E-08
positive regulation of response to stimulus	130	1.62E-05	2.15E-08	8.08E-35	NA	2.81E-05
cellular response to endogenous stimulus	85	1.37E-05	1.44E-08	4.26E-02	NA	1.29E-09
regulation of cellular protein metabolic process	115	7.39E-05	5.81E-10	1.15E-05	3.43E-02	1.14E-08
cellular response to oxygen-containing compound	44	1.93E-04	1.37E-04	7.62E-08	NA	1.03E-08
positive regulation of locomotion	17	2.79E-04	1.22E-04	6.28E-09	NA	2.12E-07

Chapter 5

EVALUATION AND CONCLUSIONS

The driving hypothesis behind this project was that some of the genes involved in the maintenance of metastatic dormancy are expressed by cells which comprise the primary tumor. Comparing tumor expression profiles from patients with late metastatic disease to such profiles from patients who appeared to remain dormant, was expected to reveal key genes that might have been associated with resumed proliferation of disseminated tumor cells. The approach taken here was to compare expression profiles between two patient classes defined on distal metastasis free survival outcomes.

Making direct, pairwise comparisons between two classes of sample, for example, "treatment vs control", is a relatively common approach, having the advantage of simple implementation and ease of interpretability. The GGI, an earlier breast cancer gene expression signature shown to be prognostic for early relapse, was developed by comparing expression profiles from Elston-Ellis Grade 1 tumors to profiles from Elston-Ellis Grade 3 tumors [25]. The GGI authors noted that patients with Grade 2 tumors presented a mixture of profiles, ranging from "Grade 1 like" to "Grade 3 like" with some intermediate responses [25]. These authors make a note of the appearance of a response "continuum" rather than a bi-modal division of patients into distinct classes [25]. This example is but one illustration of how wide the variation in human mRNA samples can be.

Unlike experimental replicates, samples from human patients cannot truly be considered identical. Biological variation in patient's life histories, exposures and genetic background can potentially influence gene expression. For example, even in normal breast tissue, there are age dependent transcriptomic changes that show striking similarity to changes that occur in breast tumors (Pirone, 2012) [50]. In the Loi A data set analyzed in this project, the median age of patients in the "Late" group was 61 vs 55 in the "Never" group. All four patients from "Loi A" in the "Never" group were negative for lymph node metastasis, whereas patients 6 of 15 "Late" group patients from "Loi A" were lymph node positive. Median tumor size in the "Loi A Late" Group was 2.4 cm vs 0.6 cm for corresponding "Never" patients. While none of these differences were statistically significant, the heterogeneity is still noteworthy. This is especially notable when the "Loi A" patients are compared to the "Loi P" patients where: median ages were 63 ("Late") and 59 ("Never"), node involvement affected 10 of 14 "Loi P Late" patients and 20 of 26 "Loi P Never" patients, and median tumor sizes were 2.0 ("Late") and 2.5 ("Never"). Unfortunately, equivalently detailed clinical data was not available for the Zhang data set. For the Symmans data set, nodal status data was available, and no significant differences between groups were observed.

Perhaps the most important difference between the "Loi A" data set and the other three is the fact that none of the "Never" group patients in "Loi A" received Tamoxifen therapy, whereas in other data sets all patients were treated with tamoxifen. Tamoxifen is an anti-estrogen drug that reduces the risk of relapse in patients with ER+ tumors [51]. The current standard for adjuvant Tamoxifen therapy, originally

based on the 1987 “B-14” trial, is 5 years of treatment [52]. In the “ATLAS”, a more recent longitudinal study, the five-year standard Tamoxifen therapy was associated with a 25.1% risk of relapse and 15% risk of breast cancer related mortality [52]. For the data evaluated in this study, the administration of tamoxifen would have influenced relapse free survival and thus which patients ended up being assigned to the “Never” group vs. the “Late” group. For example, it may have been the case for some patients that, without Tamoxifen, they would have experienced a distal relapse during the follow-up interval, and thus the expression profile from their primary tumor evaluated as a “Late” profile instead of a “Never”. In a review from 2006, Tinker et. al. specifically note that such imbalances in patient’s therapeutic histories can be a confounding factor in attempts to mine gene signatures from microarray data [53]. They go on to note various confounding factors and potential sources of error such as technical variation in sample processing, data overfitting, and of particular importance to this project, sample size considerations [53].

In one study designed to assess the effects of sample size on signature development, muscle biopsies from 69 male and 65 female human subjects were profiled for gene expression via microarray. These muscle biopsy profiles were used to develop a classifier to predict a person’s sex [54]. That analysis clearly illustrated how increased sample sizes improved both the prediction accuracy of classifiers and the number of genes with statistically significant differential expression in pairwise comparisons [54]. Notably, for a sample size of 5 per group, they found no genes with statistically significant differential expression. For a sample size of 15 per group, Stretch et al. observed fewer than 25 significant differentially expressed genes [54]. In

the context of this project, three "Never" and one "Late" group consisted of fewer than 10 patients. Muscle biopsies are a likely to be a much more homogenous source than primary tumor tissue. In light of the observed directional disagreements and weakness of adjusted *p*-values, sample size is an important factor to consider when evaluating the results of the analysis. On the other hand, there is evidence that even with smaller sample sizes, results at the pathway or process level can be reproducible across multiple separate trials [55].

With the four data sets analyzed here, there was remarkable consistency in process enrichments despite a lack of consensus at the level of differentially expressed genes. Within the set of 180 overlapping processes there were many, such as "*cell proliferation*", "*cell adhesion*", and "*MAPK Cascade*", that are potentially relevant to metastasis and dormancy; however, they are also relevant to a great many biological phenomena. One interesting case is the term "*platelet degranulation*", which was enriched in the gene lists from the "Loi A", "Loi P" and "Zhang" datasets, as well as 3 of 1000 random gene lists. This process, through which platelets release the contents of storage granules in response to injury, and is indirectly related to the proposed uPAR mediated dormancy reactivation pathway [13, 14]. Recent evidence suggests an important role for platelets in the establishment of distal metastases [56]. Among the genes annotated with the term "*platelet degranulation"* is SERPINA1, which was observed to be upregulated in "Never" patients from "Loi A", "Loi P" and "Zhang". What makes this an interesting example, is that high levels of SERPINA1 expression are associated with improved survival in ER+ breast cancer [57].

Another interesting case is the term "*ERK1 and ERK2 cascade*", which was enriched in both "Loi" data sets, though not in "Symmans" or "Zhang". This process has been implicated in the escape from a dormant state [13, 14]. However, for the two data sets where this process was enriched, the four genes differentially expressed in both are downregulated in "Never" patients from "Loi A" and upregulated in "Never" patients from "Loi P". Contradictions such as this were observed routinely in this analysis and make deeper levels of evaluation difficult.

In evaluating future opportunities, sample size and data availability appear to be key considerations. The results from the Symmans data set in Never vs Late comparisons were weak in comparison to the other three data sets. Given the issues observed with quality control probes it would be recommended to drop the Symmans data set from further analyses.

The variability observed in this analysis may result from small sample sizes, differences in the therapy patients received, or in even in the qualities of the primary tumor. It may be possible to increase sample size for the "Never" group by choosing a different cut-off between "Late" and "Never" patients, for example 10 or 15 years, across all data sets. There may also be more microarray data available in the public domain that was missed during the initial search for studies.

One data set (GEO accession: GSE7390), was excluded because its study-specific cutoff of 19.7 years would have eliminated all but 1 ER positive patient from the "Never" group. If a fixed 15-year cutoff had been used, there would have been 26

"Never" and 10 "Late" patients. Another way to achieve increased sample size would be to combine samples from multiple studies. This would require more sophisticated normalization techniques to eliminate batch effects due to variation between laboratories [29]. In addition, at least one RNA-seq based study (GEO accession: GSE119937) was available as of September 2018. As the technology matures, more clinical studies with long follow-up intervals may become available.

In addition to contradictory expression profiles, the lack of statistical significance when *p*-values were adjusted for false discovery (Benjamini-Hochberg) makes it impossible to identify a key gene set using the data evaluated in this study. Nevertheless, for three out of the four studies evaluated, enrichment results were highly inconsistent with those obtained by randomly generated lists of genes. Of equal importance is the process level reproducibility between the three data sets. Thus, it is reasonable to conclude that some of the observations presented herein are the result of underlying biological differences between the "Never" and "Late" patient classes.

In conclusion, enriched biological processes associated with proliferation, immune surveillance and hematological processes were identified. These are all implicated in the escape from dormancy. However, due to variability between data sets it remains challenging to drill down into these processes to identify a signature, or individual genes that are fundamentally involved.

REFERENCES

1. Li CI, Anderson BO, Daling JR, Moe RE. Trends in incidence rates of invasive lobular and ductal breast carcinoma. JAMA. 2003;289(11):1421-4.

2. Dai X, Xiang L, Li T, Bai Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. J Cancer. 2016;7(10):1281-94.

3. Olsnes S, Pihl A. Clinical significance of estrogen receptors in human breast cancer. Biomedicine. 1974;20(6):377-83.

4. Solomayer EF, Diel IJ, Meyberg GC, Gollan C, Bastert G. Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis. Breast Cancer Res Treat. 2000;59(3):271-8.

5. Pantel K, Brakenhoff RH. Dissecting the metastatic cascade. Nat Rev Cancer. 2004;4(6):448-56.

6. Talmadge JE, Fidler IJ. AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer Res. 2010;70(14):5649-69.

7. Paget SR, The Distribution of Secondary Growths in Cancers of the Breast. Lancet March 23 1889; 133(3421):571-3

8. Patanaphan V, Salazar OM, Risco R. Breast cancer: metastatic patterns and their prognosis. South Med J. 1988;81(9):1109-12.

9. Hess KR, Varadhachary GR, Taylor SH, et al. Metastatic patterns in adenocarcinoma. Cancer. 2006;106(7):1624-33.

10. Pan H, Gray R, Braybrooke J, et al. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. N Engl J Med. 2017;377(19):1836-1846.

11. Demicheli R, Retsky MW, Hrushesky WJ, Baum M. Tumor dormancy and surgery-driven interruption of dormancy in breast cancer: learning from failures. Nat Clin Pract Oncol. 2007;4(12):699-710.

12. Wells A, Griffith L, Wells JZ, Taylor DP. The dormancy dilemma: quiescence versus balanced proliferation. Cancer Res. 2013;73(13):3811-6.

13. Páez D, Labonte MJ, Bohanes P, et al. Cancer dormancy: a model of early dissemination and late cancer recurrence. Clin Cancer Res. 2012;18(3):645-53

14. Dittmer J. Mechanisms governing metastatic dormancy in breast cancer. Semin Cancer Biol. 2017;44:72-82.

15. Clarke B. Normal bone anatomy and physiology. Clin J Am Soc Nephrol. 2008;3 Suppl 3:S131-9.

16. Dykes SS, Hughes VS, Wiggins JM, Fasanya HO, Tanaka M, Siemann D. Stromal cells in breast cancer as a potential therapeutic target. Oncotarget. 2018;9(34):23761-23779.

17. Awolaran O, Brooks SA, Lavender V. Breast cancer osteomimicry and its role in bone specific metastasis; an integrative, systematic review of preclinical evidence. Breast. 2016;30:156-171.

18. Sasser AK, Mundy BL, Smith KM, et al. Human bone marrow stromal cells enhance breast cancer cell growth rates in a cell line-dependent manner when evaluated in 3D tumor environments. Cancer Lett. 2007;254(2):255-64.

19. Zhang XH, Jin X, Malladi S, et al. Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. Cell. 2013;154(5):1060-1073.

20. Gupta GP, Perk J, Acharyya S, et al. ID genes mediate tumor reinitiation during breast cancer lung metastasis. Proc Natl Acad Sci USA. 2007;104(49):19506-11.

21. Tseng OL, Spinelli JJ, Gotay CC, Ho WY, Mcbride ML, Dawes MG. Aromatase inhibitors are associated with a higher fracture risk than tamoxifen: a systematic review and meta-analysis. Ther Adv Musculoskelet Dis. 2018;10(4):71-90.

22. Wazir U, Mokbel L, Wazir A, Mokbel K. Optimizing adjuvant endocrine therapy for early ER+ breast cancer: An update for surgeons. Am J Surg. 2018; Article ahead of Print

23. Sestak I, Cuzick J. Markers for the identification of late breast cancer recurrence. Breast Cancer Res. 2015;17:10.

24. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671-9.

25. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006;98(4):262-72.

26. Loi S, Haibe-kains B, Desmedt C, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. J Clin Oncol. 2007;25(10):1239-46.

27. Loi S, Haibe-kains B, Desmedt C, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics. 2008;9:239.

28. Symmans WF, Hatzis C, Sotiriou C, et al. Genomic index of sensitivity to endocrine therapy for breast cancer. J Clin Oncol. 2010;28(27):4111-9.

29. Cheng Q, Chang JT, Gwin WR, et al. A signature of epithelial-mesenchymal plasticity and stromal activation in primary tumor modulates late recurrence in breast cancer independent of disease subtype. Breast Cancer Res. 2014;16(4):407.

30. Zhang Y, Sieuwerts AM, Mcgreevy M, et al. The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Breast Cancer Res Treat. 2009;116(2):303-9.

31. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7.

32. Ma XJ, Salunga R, Dahiya S, et al. A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer. Clin Cancer Res. 2008;14(9):2601-8.

33. Mosly D, Turnbull A, Sims A, Ward C, Langdon S. Predictive markers of endocrine response in breast cancer. World J Exp Med. 2018;8(1):1-7.

34. Kim RS, Avivar-valderas A, Estrada Y, et al. Dormancy signatures and metastasis in estrogen receptor positive and negative breast cancer. PLoS ONE. 2012;7(4):e35569.

35. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003;4(10):R70.

36. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

37. Mittempergher L, Saghatchian M, Wolf DM, et al. A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. Mol Oncol. 2013;7(5):987-99.

38. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5.

39. Marko NF, Frank B, Quackenbush J, Lee NH. A robust method for the amplification of RNA in the sense orientation. BMC Genomics. 2005;6:27.

40. [1] Applied BioSystems: GeneChipTM 3' IVT PLUS Reagent Kit Manual Target Preparation for GeneChipTM 3' Expression Arrays User Guide (© 2017 Thermo Fisher Scientific Inc.) P/N 703210

41. [1] Affymetrix Technical Note: Array Design for the GeneChip® Human Genome U133 Set. 2001 Part number 701133, Revision 2

42. [1] Affymetrix Technical Note: Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. 2003 Part number 701483, Revision 2.

43. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.

44. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25(3):415-6.

45. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

---

[1] Microarray technical literature (References 39 through 41) was retrieved from www.affymetrix.com on October 24, 2018. Affymetrix is now a subsidiary of Thermo Fisher Scientific Inc. and many former Affymetrix products are being marketed under the brand name 'Applied BioSystems'

46. Fresno C, Fernández EA. RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics. 2013;29(21):2810-1.

47. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE. 2011;6(7):e21800

48. Liu G, Loraine AE, Shigeta R, et al. NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res. 2003;31(1):82-6.

49. Jaccard P. The Distribution of the Flora in the Alpine Zone. New Phytol. Feb 29, 1901;11(4):37-50.

50. Pirone JR, D'arcy M, Stewart DA, et al. Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. Cancer Epidemiol Biomarkers Prev. 2012;21(10):1735-44.

51. Hubay CA, Gordon NH, Pearson OH, Marshall JS, Mcguire WL. Eight-year follow-up of adjuvant therapy for stage II breast cancer. World J Surg. 1985;9(5):738-49.

52. Smith GL. The Long and Short of Tamoxifen Therapy: A Review of the ATLAS Trial. J Adv Pract Oncol. 2014;5(1):57-60.

53. Tinker AV, Boussioutas A, Bowtell DD. The challenges of gene expression microarrays for the study of human cancer. Cancer Cell. 2006;9(5):333-9.

54. Stretch C, Khan S, Asgarian N, et al. Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. PLoS ONE. 2013;8(6):e65380.

55. Bammler T, Beyer RP, Bhattacharya S, et al. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods. 2005;2(5):351-6.

56. Leblanc R, Peyruchaud O. Metastasis: new functional implications of platelets and megakaryocytes. Blood. 2016;128(1):24-31.

57. Chan HJ, Li H, Liu Z, Yuan YC, Mortimer J, Chen S. SERPINA1 is a direct estrogen receptor target gene and a predictor of survival in breast cancer patients. Oncotarget. 2015;6(28):25815-27.

Appendix A

SCREENSHOT OF GEO SUMMARY PAGE

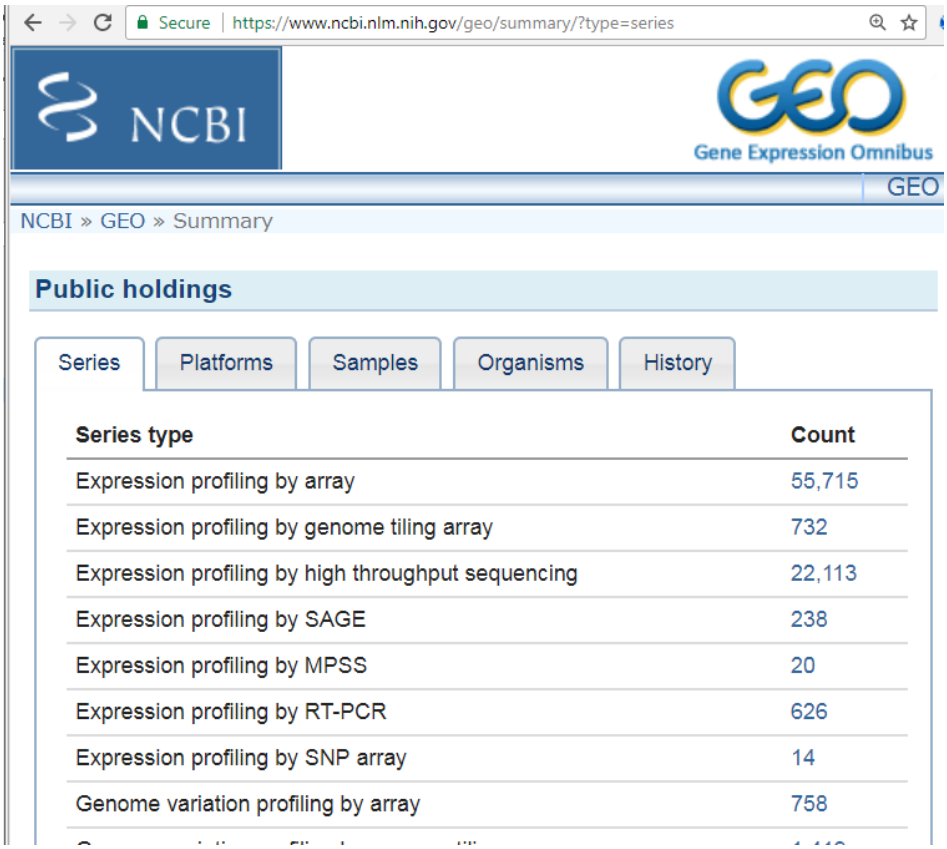

Figure 10 Screenshot of the GEO Summary page.
The url: "https://www.ncbi.nlm.nih.gov/geo/summary/?type=series" was accessed on November 11, 2018.

Appendix B

PIPELINE OVERVIEW

The analysis pipeline used here was divided into several R scripts that performed the following functions:

1. Data Retrieval, identify patient subsets, quality metrics and normalization
    i. Individual processing scripts: LoiA_Analysis.r, LoiP_Analysis.r, Zhang_Analysis.r, Symmans_Analysis.r
    ii. Parsing Functions: KeyValueExtract.r
2. Differential expression analysis, and DAVID Query Submission
    i. Limma_DAVID_Analysis.r
3. Evaluate overlapping ontology
    i. Ontology_Overlap.r

For Randomized Analyses:

1. Generate 1000 Random Probe Lists, divided into batches of 100
    i. David_Random_Genelists.r
2. Submit lists as DAVID queries and retrieve results.
    i. David_Query.r
3. Tabulate query results and generate histograms.
    i. randomGene_analaysis.r

All of the pipeline related scripts, and accompanying documentation have been deposited in a repository on GitHub URL:

https://github.com/afaranda/NeverLatePipeline

Appendix C

PERSONAL COMMUNICATION WITH NIH STAFF RE: GOTERM_BP_FAT

DAVID Ontology category FAT Inbox x

Adam Pater-Faranda <abf@udel.edu> Thu, Mar 1, 9:33 PM

to bsherman

Greetings,

I'd like to learn more about the criteria that DAVID uses to include terms in the category 'GOTERM_BP_FAT'. Is there a literature reference where this category is defined? when I click on the "?" button under the functional annotation clustering tool, it leads to a "404" page. Any information you could provide me with about the FAT categories would be very helpful.

Best,

Adam

Sherman, Brad (NIH) [C] <bsherman@mail.nih.gov> Fri, Mar 2, 4:25 PM

to me

Hi Adam,

Thanks you for contacting us. I apologize for the dead help link. It was pointing to our forum which we recently had to shutdown due to spam. Please see our explanation below for the GO categories in DAVID:

GOTERM_XX_ALL includes all terms in the Gene Ontology hierarchy for the given main branch, i.e. Biological Process(BP), Cellular Component(CC), and Molecular Function(MF). GOTERM_XX_1 and GOTERM_XX_FAT are subsets of GOTERM_XX_ALL.

We separate the GO ontology into levels based on parent and child terms in a hierarchy. Starting at level 1 (i.e. GOTERM_XX_1), terms have the broadest meaning and are therefore the least specific. For instance, "cellular process" is just below the main "Biological Process" branch.

GOTERM_XX_FAT is mostly made up of higher level terms in order to gain the more specific terms in the output and reduce some redundancy. Some groups have created subsets called GO Slim which include the broadest GO terms such as "immune system process" and filter out the more specific terms such as "T cell selection". We are more interested in the specific terms, so we created the GO Fat subset to filter out the broad terms and include the specific ones. Due to the structure of GO, the levels cannot universally define the specificity of a given term. Therefore, we defined the term specificity based on the number of child terms to filter out the broadest terms in the hierarchy.

The " GOTERM_XX_DIRECT" categories refer to the Gene Ontology annotation assigned directly by the source of the data that DAVID uses (i.e. NCBI, Uniprot). For the "_ALL" and "_FAT" GO categories, we expand the direct mappings to the parent, grandparent, etc terms of the directly annotated term.

Regards, Brad

Brad T. Sherman, M.S. (Contractor)

Laboratory of Human Retrovirology and Immunoinformatics

Appendix D

ARRAY QUALITY METRICS RESULTS

a)

tadata and outlier detection overview

		sampleNames	*1	*2	*3	group	geo_accession	geo_accn_hg.u133plus2	series	age	grade	size	node	DFS_TIME	EVENT_DFS
		SM150945.CEL.gz				LATE	GSM150945		OXFT	70		5	0	2266	1
☐	2	GSM151020.CEL.gz				LATE	GSM151020		KIT	57	2	1.4	1	2372	1
☐	3	GSM151028.CEL.gz				LATE	GSM151028		KIT	60	2	2.4	1	2555	1
☐	4	GSM151042.CEL.gz				LATE	GSM151042		KIT	51	2	3.3	1	2646	1
☐	5	GSM151050.CEL.gz				LATE	GSM151050		KIT	62	2	2.4	0	2372	1
☐	6	gsm65336.cel.gz				LATE	GSM65336		KIT	64	1	2.6	0	3468	1
☐	7	gsm65338.cel.gz				LATE	GSM65338		KIT	73	3	3.3	1	3256	1
☑	8	gsm65377.cel.gz	x			LATE	GSM65377		OXFT	69	1	5	1	2070	1
☐	9	gsm65378.cel.gz				LATE	GSM65378		OXFT	64	3	5.5	1	2600	1
☐	10	gsm65770.cel.gz				LATE	GSM65770		KIU	41	1	1.8	0	2493	1
☐	11	gsm65782.cel.gz				LATE	GSM65782		KIU	56	1	2.2	0	2341	1
☐	12	gsm65815.cel.gz				LATE	GSM65815		KIU	56	1	2.2	0	2158	1
☐	13	gsm65820.cel.gz				NEVER	GSM65820		OXFU	44		0	0	5305	0
☐	14	gsm65823.cel.gz				NEVER	GSM65823		OXFU	61		0.3	0	5178	0
☐	15	gsm65826.cel.gz				LATE	GSM65826		OXFU	61		1.4	0	2125	1
☐	16	gsm65827.cel.gz				NEVER	GSM65827		OXFU	58		0.9	0	5029	0
☐	17	gsm65831.cel.gz				NEVER	GSM65831		OXFU	55	3	3	0	5029	0
☐	18	gsm65834.cel.gz				LATE	GSM65834		OXFU	64	2	1.5	0	4579	1
☐	19	gsm65837.cel.gz				LATE	GSM65837		OXFU	64	3	2.1	0	1875	1

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

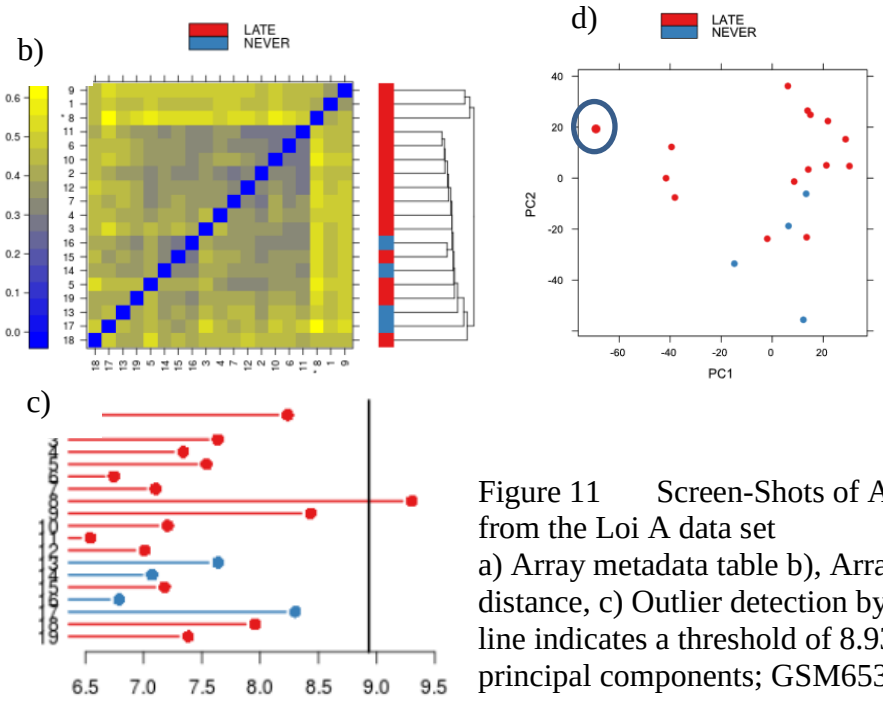

Figure 11 Screen-Shots of AQM results from the Loi A data set
a) Array metadata table b), Array clustering by distance, c) Outlier detection by distance (black line indicates a threshold of 8.93). d) First 2 principal components; GSM65377 circled.

- Array metadata and outlier detection overview

	array	sampleNames	*1	*2	*3	group	geo_accession	geo_accn_hg.u133a	geo_accn_hg.u133b	age	grade	size	pgr	node	DFS_TIME
☐	1	GSM151259.CEL.gz				LATE	GSM151259			46		3	1	0	3821
☐	2	GSM151260.CEL.gz				LATE	GSM151260			71		2	1	1	3023
☐	3	GSM151261.CEL.gz				NEVER	GSM151261			58		1.3		1	6037
☐	4	GSM151262.CEL.gz				LATE	GSM151262			57	2	2	1	1	4370
☐	5	GSM151263.CEL.gz				NEVER	GSM151263			69		1.5	1	1	5313
☐	6	GSM151264.CEL.gz				LATE	GSM151264			67	1	2	0	0	2933
☐	7	GSM151265.CEL.gz				NEVER	GSM151265			56		2.5	1	1	5189
☐	8	GSM151270.CEL.gz				NEVER	GSM151270			57	2	3.5	1	1	5956
☐	9	GSM151272.CEL.gz				NEVER	GSM151272			61	2	2.5	1	1	6151
☐	10	GSM151273.CEL.gz				NEVER	GSM151273			52	2	2.5	1	1	5967
☑	11	GSM151274.CEL.gz		x		NEVER	GSM151274			63	2	1.5	1	1	5737
☐	12	GSM151276.CEL.gz				NEVER	GSM151276			58	1	2.5	1	1	5820
☐	13	GSM151278.CEL.gz				NEVER	GSM151278			60	2	1.5	1	1	5959
☐	14	GSM151279.CEL.gz				LATE	GSM151279			58		3.5	1	1	3685
☐	15	GSM151280.CEL.gz	x	x		NEVER	GSM151280			60	2	1.5	1	1	5725
☐	16	GSM151281.CEL.gz				LATE	GSM151281			66	2	2.5	1	1	2622
☐	17	GSM151282.CEL.gz				NEVER	GSM151282			51	3	3.5	1	1	5554
☐	18	GSM151283.CEL.gz				NEVER	GSM151283			62		3	0	0	5527
☐	19	GSM151284.CEL.gz				NEVER	GSM151284			53	2	2	0	1	5593
☐	20	GSM151285.CEL.gz				LATE	GSM151285			62		2	1	1	3355
☐	21	GSM151287.CEL.gz				NEVER	GSM151287			59	3	2	0	1	5518
☐	22	GSM151289.CEL.gz				NEVER	GSM151289			62	2	3	1	1	5189
☐	23	GSM151291.CEL.gz				LATE	GSM151291			56	2	2	0	0	4774
☐	24	GSM151292.CEL.gz				NEVER	GSM151292			53		4	1	0	5132
☐	25	GSM151293.CEL.gz				NEVER	GSM151293			59		1.1	1	1	5274
☑	26	GSM151294.CEL.gz		x		NEVER	GSM151294			58	1	3	1	1	5311
☐	27	GSM151298.CEL.gz				LATE	GSM151298			63	3	3	0	1	1923
☐	28	GSM151299.CEL.gz				LATE	GSM151299			61		1.5	1	1	4465
☐	29	GSM151300.CEL.gz				NEVER	GSM151300			66	2	2	1	0	5214
☐	30	GSM151305.CEL.gz				NEVER	GSM151305			61	1	2	1	1	5189
☐	31	GSM151306.CEL.gz				NEVER	GSM151306			67	1	1.9	1	0	5054
☐	32	GSM151309.CEL.gz				NEVER	GSM151309			50	3	3	1	1	4973
☐	33	GSM151310.CEL.gz				LATE	GSM151310			63	3	3	1	1	3795
☐	34	GSM151313.CEL.gz				NEVER	GSM151313			70	2	2.5	1	1	4988
☐	35	GSM151322.CEL.gz				LATE	GSM151322			65		3	1	1	3304
☐	36	GSM151323.CEL.gz				LATE	GSM151323			63	2	1.5	1	0	3004
☐	37	GSM151342.CEL.gz				LATE	GSM151342			70	2	1.6	1	1	3751

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

Figure 12 Screen shot of Array Quality Metrics Metadata Overview from the Loi P Dataset

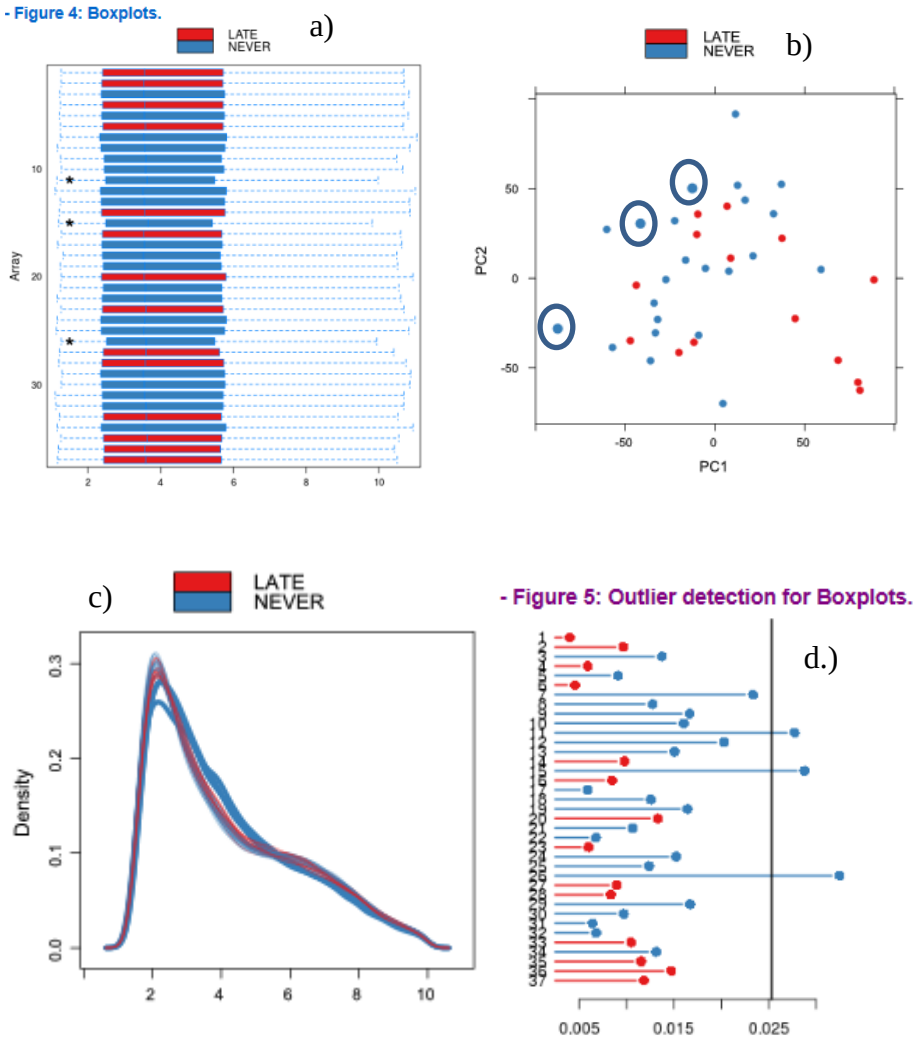

Figure 13 Screen shot of Array Quality Metrics Results from the Loi P dataset a) Box plots of probe intensities for each array, b) First two principal components (outliers circled) c) Density plots illustrating probe intensity distributions d) Outlier detection based on boxplots (black line is threshold of K_a = 0.0253 Kolmogorv-Smirnov vs. pooled density)

a)

- Array metadata and outlier detection overview

	array	sampleNames	*1	*2	*3	group	geo_accession	DMFS_TIME	EVENT_DMFS	Patient_ID	supplementary_file	ScanDate
☐	1	GSM305151.CEL.gz				LATE	GSM305151	64.9	1	3251	GSM305151.CEL.gz	05/12/05 09:26:04
☐	2	GSM305172.CEL.gz				LATE	GSM305172	75.9	1	3275	GSM305172.CEL.gz	05/12/05 15:10:51
☐	3	GSM305194.CEL.gz				LATE	GSM305194	104.1	1	2536	GSM305194.CEL.gz	10/11/06 09:24:34
☐	4	GSM305197.CEL.gz				NEVER	GSM305197	192.6	0	2647	GSM305197.CEL.gz	10/11/06 09:40:43
☐	5	GSM305198.CEL.gz				LATE	GSM305198	108.6	1	2494	GSM305198.CEL.gz	10/11/06 09:45:56
☐	6	GSM305199.CEL.gz				NEVER	GSM305199	191.9	0	2733	GSM305199.CEL.gz	10/11/06 09:51:13
☐	7	GSM305200.CEL.gz				LATE	GSM305200	139.5	1	2639	GSM305200.CEL.gz	10/11/06 09:21:22
☐	8	GSM305203.CEL.gz				LATE	GSM305203	164.3	1	2308	GSM305203.CEL.gz	10/11/06 09:37:16
☐	9	GSM305211.CEL.gz				LATE	GSM305211	69.2	1	21790	GSM305211.CEL.gz	10/11/06 11:44:06
☐	10	GSM305212.CEL.gz				NEVER	GSM305212	191.3	0	21148	GSM305212.CEL.gz	10/11/06 11:49:26
☐	11	GSM305215.CEL.gz				NEVER	GSM305215	177.3	0	2548	GSM305215.CEL.gz	10/11/06 11:26:59
☐	12	GSM305225.CEL.gz				NEVER	GSM305225	182.1	0	2392	GSM305225.CEL.gz	10/11/06 12:28:01
☐	13	GSM305226.CEL.gz				NEVER	GSM305226	176.3	0	2377	GSM305226.CEL.gz	10/11/06 12:33:10
☑	14	GSM305264.CEL.gz	x	x		NEVER	GSM305264	179.6	0	77	GSM305264.CEL.gz	11/11/05 11:02:45

- Figure 1: Distances between arrays.

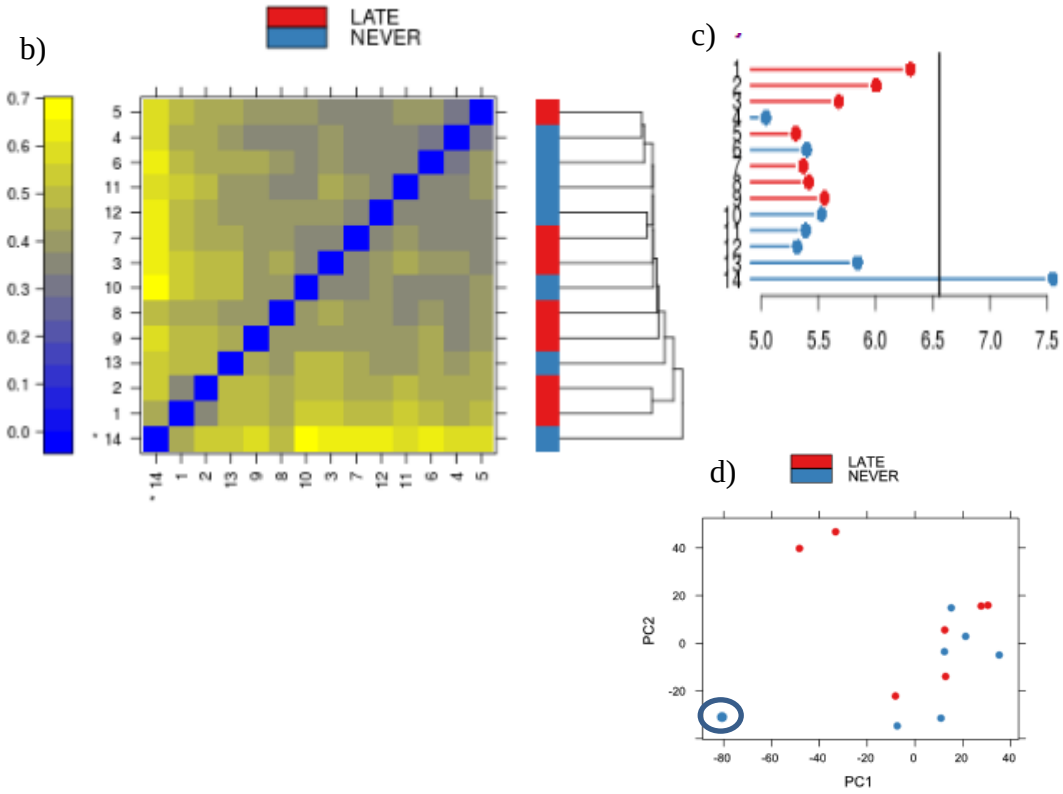

Figure 14 Screen Shot of Array Quality Metrics Results from the Zhang dataset a) metadata overview, b) Array clustering by distance, c) Outlier detection by distance (black line indicates an outlier threshold of 6.55), d) First two principal components (outlier circled)

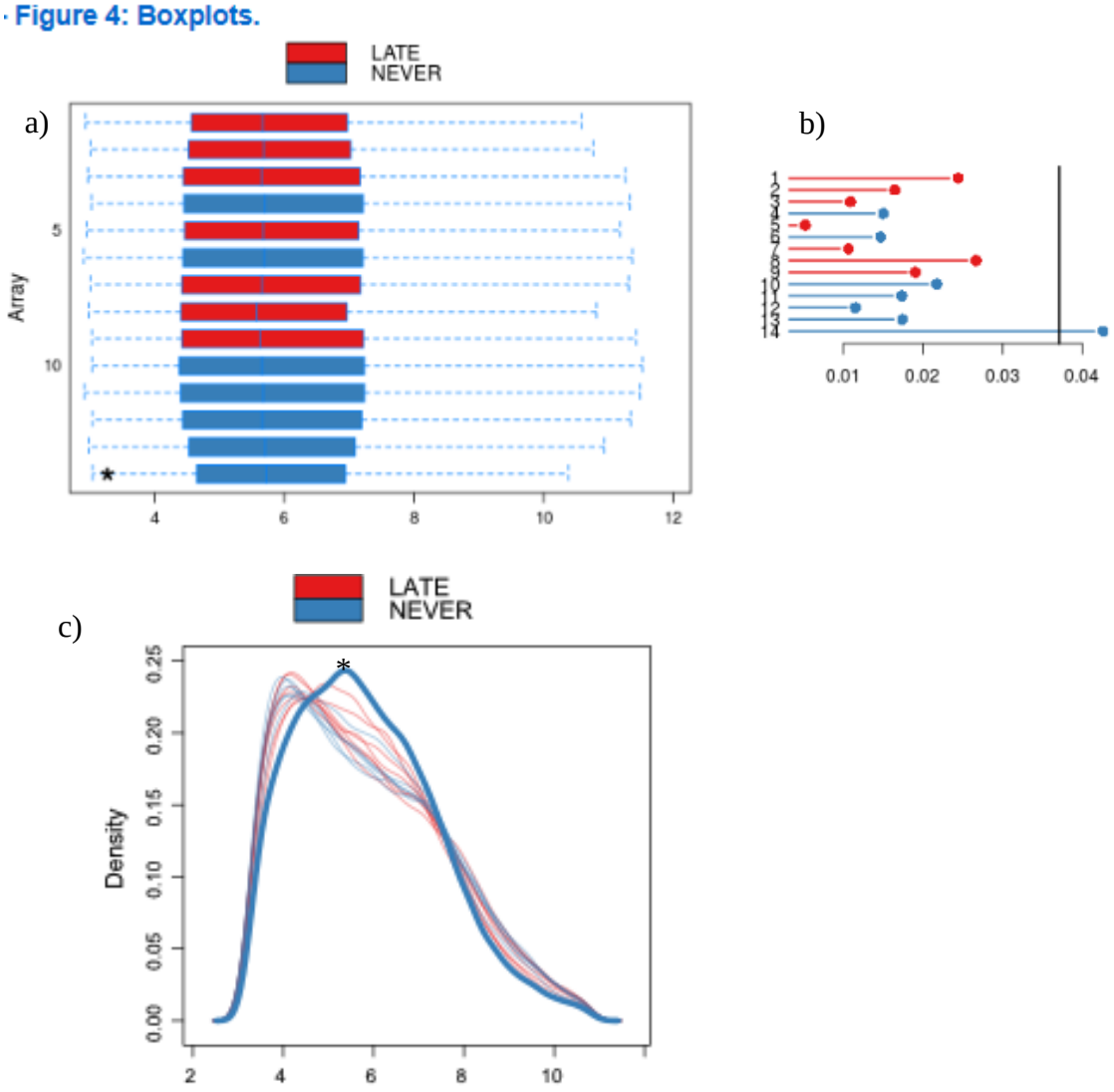

Figure 15 Screen shot of Array Quality Metrics Results from the Zhang dataset a) Box plots of probe intensities for each array (asterisk marks outlier) b) Outlier detection based on boxplots (black line is threshold of K_a = 0.0371 Kolmogorv-Smirnov vs. pooled density), c) Density plots illustrating probe intensity distributions (asterisk marks outlier).

	array	sampleNames	*1	*2	*3	group	geo_accession	profiling lab	nodal status (0=negative, 1=positive, na=not applicable)	EVENT_DMFS
☑	1	GSM441627.CEL.gz	x	x		LATE	GSM441627	JBI	0	1
☑	2	GSM441691.CEL.gz	x	x		LATE	GSM441691	JBI	1	1
☑	3	GSM441692.CEL.gz	x			LATE	GSM441692	JBI	1	1
☑	4	GSM441705.CEL.gz	x			LATE	GSM441705	JBI	1	1
☐	5	GSM441727.CEL.gz				LATE	GSM441727	MDA	1	1
☐	6	GSM441728.CEL.gz				LATE	GSM441728	MDA	1	1
☐	7	GSM441733.CEL.gz				LATE	GSM441733	MDA	1	1
☐	8	GSM441743.CEL.gz				LATE	GSM441743	MDA	0	1
☐	9	GSM441747.CEL.gz				LATE	GSM441747	MDA	1	1
☐	10	GSM441766.CEL.gz				LATE	GSM441766	MDA	0	1
☐	11	GSM441770.CEL.gz				LATE	GSM441770	MDA	0	1
☐	12	GSM441782.CEL.gz				LATE	GSM441782	MDA	1	1
☐	13	GSM441785.CEL.gz				LATE	GSM441785	MDA	1	1
☐	14	GSM441808.CEL.gz				LATE	GSM441808	MDA	1	1
☐	15	GSM441812.CEL.gz				LATE	GSM441812	MDA	0	1
☐	16	GSM441816.CEL.gz				LATE	GSM441816	MDA	0	1
☐	17	GSM441818.CEL.gz				LATE	GSM441818	MDA	0	1
☐	18	GSM441823.CEL.gz				LATE	GSM441823	MDA	1	1
☐	19	GSM441826.CEL.gz				NEVER	GSM441826	MDA	1	0
☐	20	GSM441829.CEL.gz				LATE	GSM441829	MDA	1	1
☐	21	GSM441830.CEL.gz				LATE	GSM441830	MDA	1	1
☐	22	GSM441834.CEL.gz				LATE	GSM441834	MDA	0	1
☐	23	GSM441836.CEL.gz				LATE	GSM441836	MDA	0	1
☐	24	GSM441838.CEL.gz				LATE	GSM441838	MDA	1	1
☐	25	GSM441840.CEL.gz				LATE	GSM441840	MDA	0	1
☐	26	GSM441850.CEL.gz				LATE	GSM441850	MDA	1	1
☐	27	GSM441868.CEL.gz				LATE	GSM441868	MDA	1	1
☐	28	GSM441874.CEL.gz				LATE	GSM441874	MDA	0	1
☐	29	GSM441884.CEL.gz				LATE	GSM441884	MDA	NA	1
☐	30	GSM441889.CEL.gz				NEVER	GSM441889	MDA	0	0
☐	31	GSM441897.CEL.gz				NEVER	GSM441897	MDA	1	0
☐	32	GSM441905.CEL.gz				NEVER	GSM441905	MDA	0	0
☐	33	GSM441906.CEL.gz				NEVER	GSM441906	MDA	0	0
☐	34	GSM441907.CEL.gz				NEVER	GSM441907	MDA	1	0
☐	35	GSM441920.CEL.gz				LATE	GSM441920	MDA	0	1
☐	36	GSM441921.CEL.gz				NEVER	GSM441921	MDA	1	0

Figure 16 Screen shot of Array Quality Metrics Results from the Symmans dataset Metadata overview

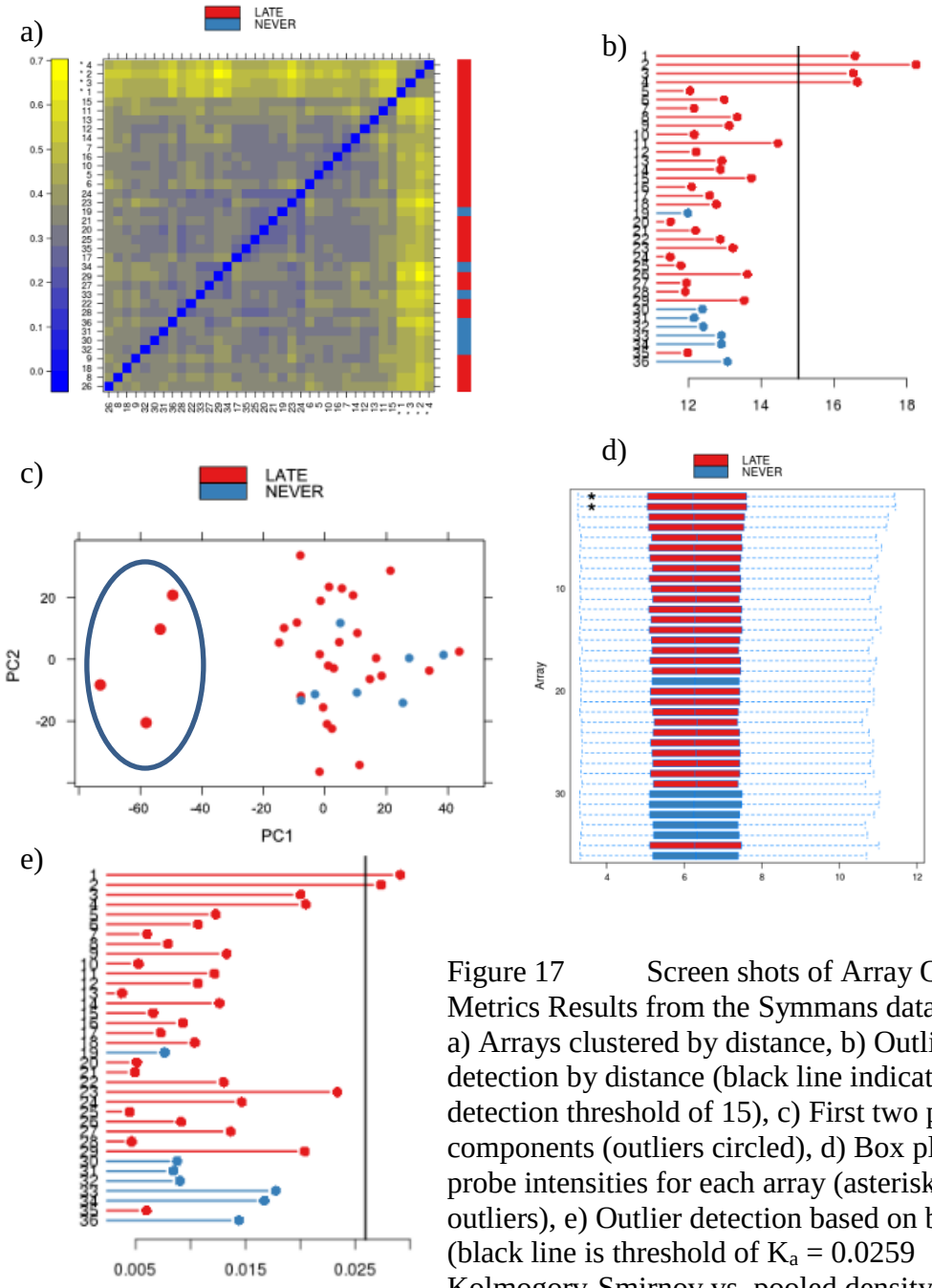

Figure 17 Screen shots of Array Quality Metrics Results from the Symmans dataset a) Arrays clustered by distance, b) Outlier detection by distance (black line indicates a detection threshold of 15), c) First two principal components (outliers circled), d) Box plots of probe intensities for each array (asterisks indicate outliers), e) Outlier detection based on boxplots (black line is threshold of $K_a = 0.0259$ Kolmogorv-Smirnov vs. pooled density)