Heterogeneity in global gene expression profiles between biopsy specimens taken peri-surgically from primary ER-positive breast carcinomas

Authors

Elena López-Knowles[1,2,†] *, Qiong Gao[2,†], Maggie Chon U Cheang[3], James Morden[3], Joel Parker[4], Lesley-Ann Martin[2], Isabel Pinhel[1,2^], Fiona McNeill[1], Margaret Hills[1], Simone Detre[1], Maria Afentakis[1], Lila Zabaglo[1], Andrew Dodson[1], Anthony Skene[5], Chris Holcombe[6], John Robertson[7], Ian Smith[1], Judith M Bliss[3], Mitch Dowsett[1,2] on behalf of the POETIC trialists.

Emails: elena.lopez-knowles@icr.ac.uk, Alice.gao@icr.ac.uk, Maggie.Cheang@icr.ac.uk, James.Morden@icr.ac.uk, parkerjs@email.unc.edu, Lesley-ann.martin@icr.ac.uk, isabelpinhel@hotmail.com, fiona.macneill@rmh.nhs.uk, Margaret.hills@icr.ac.uk, Simon.Detre@icr.ac.uk, maria.afentakis@icr.ac.uk, lila.zabaglo@icr.ac.uk, Andrew.dodson@icr.ac.uk, Anthony.skene@rbch.nhs.uk, chris.holcombe@rlbuht.nhs.uk, john.robertson@nottingham.ac.uk, ian.smith@rmh.nhs.uk, Judith.Bliss@icr.ac.uk, Mitch.dowsett@icr.ac.uk

[1] Royal Marsden Hospital, London, United Kingdom

[2] Breast Cancer Now Research Centre, The Institute of Cancer Research, London, United Kingdom

[3] Clinical Trials and Statistics Unit, The Institute of Cancer Research, London, United Kingdom

[4] UNC, Chapel Hill, North Carolina, USA

[5] Royal Bournemouth Hospital, Bournemouth, United Kingdom

[6] Royal Liverpool University Hospital, Liverpool, United Kingdom

[7] Queen's Medical Centre, Nottingham, United Kingdom

^ Current affiliation: Kingston University, London, United Kingdom

[†] Contributed equally

* Corresponding author

Abstract

Introduction

Gene expression is widely used for the characterization of breast cancers. Variability due to tissue heterogeneity or measurement error or systematic change due to peri-surgical procedures can affect measurements but is poorly documented. We studied the variability of global gene expression between core-cuts of primary ER+ breast cancers and the impact of delays to tissue stabilization due to sample x-ray and of diagnostic core-cutting.

Methods

Twenty-six paired core-cuts were taken immediately after tumour excision and up to 90 minutes delay due to sample x-ray; 57 paired core-cuts were taken at diagnosis and 2 weeks later at surgical excision. Whole genome expression analysis was conducted on extracted RNA. Correlations and differences were assessed between the expression of individual genes, gene-sets/signatures and intrinsic subtypes.

Results

Twenty-three and 56 sample pairs, respectively, were suitable for analysis. The range of correlations for both sample sets were similar with the majority being >0.97 in both. Correlations between pairs for 18 commonly studied genes were also similar between the studies and mainly with Pearson correlation coefficients >0.6 except for a small number of genes which had a narrow-dynamic range (e.g. *MKI67*, *SNAI2*). There was no systematic difference in intrinsic subtyping between the first and second sample of either set but there was c.15% discordance between the subtype assignments between the pairs, mainly where the subtyping of individual samples was less certain. Increases in the expression of several stress/early-response genes (e.g. *FOS, FOSB, JUN*) were found in both studies and confirmed findings in earlier smaller studies. Increased expression of *IL6*, *IGFBP2* and *MYC* (by 17%, 14% and 44%, respectively) occurred between the samples taken 2-weeks apart and again confirmed findings from an earlier study.

Conclusions

There is generally good correlation in gene expression between pairs of core-cuts except where genes have a narrow dynamic range. Similar correlation coefficients to the average gene expression profiles of intrinsic subtype, particularly LumA and LumB, can lead to discordances between assigned subtypes. Substantial changes in expression of early response genes occur within an hour after surgery and in *IL6, IGFB2* and *MYC* as a result of diagnostic core-cut biopsy.

Trial Registration

Trial Number CRUK/07/015. Study start date September 2008.

Keywords

Breast Cancer, Gene expression, heterogeneity

Introduction

Molecular analyses of primary breast cancer for both research and patient management are now commonplace. Measurements may be made on diagnostic core-cut biopsies or surgical excisions that frequently comprise a very small fraction of the tumour. In so-called window-of-opportunity studies patients are exposed to medical therapy between diagnosis and surgery [1] and comparisons are made between samples taken at both time points. Valid interpretation of these studies depends on knowledge of any variability or systematic changes in the respective biomarkers that occur in the absence of treatment. Variability/heterogeneity may lead to false rejection of a true effect while systematic differences between diagnostic and surgical specimens may lead to artifactual changes being falsely ascribed to an intervention. For example, we have previously described the highly significant impact of specimen type (core-cut vs excision) on pAKT and pERK1/2 staining [2]. Pre-treatment/post treatment comparison of biomarkers might also be affected by the taking of the diagnostic biopsy and changes due to cold ischemia between resection and tissue stabilization/fixation.

The effect of cold ischemia time has been studied in small cohorts of breast cancer with up to 24 hours elapsed time before fixation, snap freezing or placement in RNA later [3-5]. No studies have directly examined the impact of the short time delay (20-60 minutes) resulting from sending specimens for x-ray, a frequent practise during breast cancer surgery to ensure the removal of the lesion (e.g. non-palpable mass, calcifications) and/or to check for adequate surgical margins, even in clinically palpable tumours. A small number of studies have evaluated gene expression changes over a longer period of time between biopsies [6-8]. For example, Jeselsohn identified 14 genes, including 9 immune-related that differed between core cuts and excision taken from 21 patients 6-65 days apart (mean 30 days).

Our primary objectives were to use genome-wide expression profiling to determine more comprehensively the variability and systematic changes in the expression of genes or pre-specified genesets or subtype classifications (i) between two core biopsies taken (A) immediately after excision and (B) after sample x-ray and (ii) between diagnostic core biopsies (D) and surgical core biopsies (S) two weeks later in the absence of any intervention.

Patients and Methods

Patients and tissues

Study I. To answer the first objective we accessed tissues collated from a previously published study [2]. Core cut biopsies (14-gauge needle) were taken from 26 surgical specimens and placed in RNAlater immediately after resection (sample A) and again after X-ray of the excised tumour (sample B). The time elapsed between samples A and B was recorded in the surgical report form.

Study II. To answer the second objective we accessed tissues from the no-treatment arm of The PeriOperative Endocrine Therapy - Individualising Care (POETIC) trial that randomized post-menopausal patients with primary ER+ breast cancer from 120 UK centres (2:1) to receive two weeks' non-steroidal aromatase inhibitor (AI) or no-treatment for two weeks prior to surgery[1, 9].

At least 1 RNA later stored sample was available from 33.5% (1493/4456) of patients or paired from 13.2% (589/4456) of patients of the poetic trial. 227 control samples were subjected to RNA extraction. Expression analyses were conducted when a pair of RNA extracts was available with RIN >4. This amounted to 57 pairs of samples from control patients taken at diagnosis (D) and surgery (S).

Ethics statement

Patient consent and ethics approval for the collection and analysis of breast cancer tissue samples was provided by the Royal Marsden Hospital for Study I. Ethical approval for POETIC (Trial Number CRUK/07/015) was provided by NRES Committee London –South East.

Gene expression analysis, data pre-processing, data analyses and statistical methods.

The detailed methodology is described in the supplementary information.

In brief, extracted RNA was amplified, labeled and hybridized on Illumina global gene expression BeadChips. Illumina raw data was extracted using GenomeStudio software and transformed, normalized and batch-corrected. Paired samples were excluded from further analysis if their fraction of detected genes was <30% and probes were filtered out if they were not detected in any sample. Gene expression data from this study is deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73237) with accession number GSE73237 [10].

Entrez Gene ID was used as gene identifier in gene signatures. The HumanHT-12_V4_0_R2_15002873_B annotation file was used to map the EntrezGeneIDs to the corresponding Illumina probe IDs. Gene signature scores were weighted averages.

We evaluated three candidate gene sets: i) metagene wound healing signature [11]; ii) immune response metagene [12] and iii) 13 of the 14 genes identified as changing in the Jeselsohn study [6] (SNAI1 was not detected on the Illumina platform). We also studied the effects on 18 pre-specified genes that we selected as being particularly relevant to breast cancer from prior studies.

Each tumour sample was classified into one of the five intrinsic subtypes based on the PAM50 classifier as described in the supplementary information.

Pearson and Spearman correlations were used to assess the associations. Univariate paired or unpaired T-tests together with multivariate permutation tests were used to identify differentially expressed genes between the paired samples. The significantly differentially expressed genes were subjected to Ingenuity Pathway Analysis (IPA). The significance of the difference between 2 correlation coefficients obtained in study I and study II respectively was calculated using the Fisher r-to-z transformation [13]. GraphPad Prism 6 (Graphpad Software Inc.) was used for some of the statistical analyses in this study.

Results

Study I

Sufficient RNA was available from 26 sample pairs with up to 90 minutes between samples A and B. Three pairs were excluded due to low fraction of detected genes, leaving 23 pairs with a time interval of 20 to 60 minutes (median 30) for downstream data analysis. Patient demographics are described in Table S1.

Variability in gene expression between samples

On hierarchical clustering 16 (70%) of the pairs clustered together (Figure 1A). The correlation of the gene expression for the 24,395 probes between samples A and B provides an overall assessment of the similarity of transcriptional profiles between the samples. The Pearson correlation coefficient r values ranged from 0.91 to >0.99 (Figure S1). Nine selected pairs in Figure S2 represent the range of variability: 3 sets of 3 pairs with a coefficient >0.99, 0.98 or 0.91-0.94. Correlation was also determined between paired expression levels of 18 pre-selected genes frequently reported in breast cancer (Table S2, Figure S2). The correlation was above >0.6 and highly statistically significant for all genes, except for *MKI67* (r=0.35, p=0.10), *SNAI2* (r=0.43, p=0.04) and *PGR*

(r=0.52, p=0.01) (Table 1). Upload of the full data set to GSE73237 [10] allows investigators to assess the correlation/variability of their genes of interest.

Effect of time to fixation on gene expression

Using class comparison method with False Discovery Rate (FDR) <5% no significant systematic differences in expression were found between samples A and B. However, 68 genes had a p<0.005 and fold-change ≥1.25 (19 upregulated and 49 downregulated). Table 2 shows the top 8 of these genes ordered according to fold-change. The genes included early response (*RGS1*, *RGS2*), mitochondrial ATP synthase (*ATP5C1*) and stress response genes (*DUSP1*, *FOSB*). Ingenuity Pathway Analysis (IPA) of the 68 genes using Benjamini-Hochberg multiple testing corrected B-H p-value <0.05, identified 6 canonical pathways (Table S3A). These were mainly associated with metabolism or signalling, the most significant being oxidative phosphorylation (B-H p-value <0.005) and mitochondrial dysfunction (B-H p-value <0.005). The top networks identified also included metabolism (Table S3B).

Change in expression of 116 genes correlated with time elapsed at p<0.005 (Table S4) but none were significant by their adjusted p-value. IPA of the 116 genes identified 28 pathways that were significantly changed at p<0.05. The most significant were adipogenesis and mitochondrial dysfunction and the main networks were inflammation and metabolic disease (Tables S5 and S6). There were only 2 genes in common between the 68 (paired differences) and 116 (time elapsed) gene lists (*SCD* and *AGPAT2* involved in fatty acid biosynthesis).

Two of the 18 genes pre-selected as frequently reported showed a modest but statistically significant difference between samples A and B: BAG1 (mean 3% decrease, p=0.026), MAPT (mean 19% decrease, p=0.007) (Table 1).

Analysis of candidate gene signatures and subtypes

There were no significant differences in the Wound Healing signature score[11] or an immune-response metagene [12]. One of the 13 genes identified to be changing in the

Jeselsohn study (*IL6*) showed an 11% increase (Wilcoxon matched-pairs signed rank test: p=0.014) between samples A and B [6].

Concordance for intrinsic subtypes between the sample pairs is shown in Table S7. The majority of these ER+ samples were Luminal, as expected. Three tumours showed discordance between samples at timepoint A and timepoint B: two Luminal A samples at time point A were scored as Luminal B or normal at time point B; one luminal B at time point A was rated as Luminal A at time point B. For each tumour, we calculated the numerical differences in the correlation coefficients to each of the LumA, LumB, and HER2-enriched centroids for each of samples A and B. As demonstrated in Figure S4A, these 3 cases with discordant intrinsic subtypes between the time points A and B had the median values of numeric difference between their LumA and LumB centroid correlations of 0.08 and 0.32 when compared with a median difference of 0.54 (95% C.I. 0.17-0.61) and 0.52 (95% C.I. 0.10-0.54) for the concordant samples at time points A and B respectively.

Study II

From the 57 pairs, 56 passed microarray QC analysis. Patient demographics are described in Table S1.

Variability in gene expression between samples

Seventy-three percent (41/56) of pairs clustered together on hierarchical clustering (Figure 1B). The correlation of the gene expression for the 32,332 probes between the 2 samples ranged from 0.86 to >0.99 with a median correlation of 0.97 (Figure S5). As in study I, we evaluated the Pearson correlation coefficients between paired expression levels on 18 selected genes (Table S2, Figure S6). The correlation was above >0.6 except for *SNAI2* (r=0.48), *MKI67* (r=0.52), and *GPR160* (r=0.55).

Gene expression comparison between baseline and surgery core

Thirty-nine genes (44 probes) were differentially expressed between biopsies D and S at FDR<5% and fold-change > 1.25. The 39 genes included 11 early response genes (*FOS*, *JUN*, *RGS1*), 6 stress response/immune genes (*DUSP1*, *GADD45B*, *ATF3*), 4 snoRNA (*SNORD3C*, *SNORD3D*), 4 haemoglobin (*HBA2*, *HBB*) and 5 genes associated to breast cancer progression (*SIK1*, *TOB1*, *BHLHB2*). Table 2B shows the top 8 genes identified. IPA analysis of the 39 genes identified 76 pathways affected (B-H p-value <0.05) (Table S8). Sixty per cent of the pathways identified were due solely to *FOS* and *JUN*. The most common enriched networks were proliferation and metabolism (Table S9). None of the 18 pre-selected genes showed a statistically significant change between samples D and S (Table 1).

Analysis of candidate gene signatures and subtypes

There were no significant differences in the Wound Healing signature [11] or the immune response gene signature [12] between samples D and S. Of the 14 detected significantly differ genes described by Jeselsohn, two immune-related genes (*IL6* and *IGFPB2*) and one other gene (*MYC*) were significantly increased in their expression in sample S by 17%, 14%, and 44%, respectively. The changes in IL6, IGFBP2 and MYC did not significantly correlate with one another.

Most samples were Luminal (Table S7B). Six of 39 (15%) tumours classified as Luminal A at baseline were classified as Luminal B at surgery, and four of 14 tumours classified as Luminal B at baseline were classified as Luminal A at surgery (29%, 4/14). Among the 14 cases with discordant intrinsic subtypes between the baseline and surgery, the median values of numeric difference between their Luminal A and Luminal B centroid correlations were 0.089 (95% C.I. 0.02-0.49) and 0.031 (95% C.I. 0.12-0.34) when compared with median values of 0.50 (95% C.I. 0.26-0.55) and 0.50 (95% C.I. 0.26-0.53) for the concordant samples at baseline and surgery respectively (Figure S4B). Interestingly, the one LumB/HER2-E subtype discordant case also had <0.3 between the LumB/HER2-E centroids.

Study I and Study II common genes

Nine of the top 20 genes significantly different with FDR <5% and p<0.005 between samples D and S in study II were also significant with a p<0.05 between samples A and B in study I (Table S10). These included *FOS*, *JUN* and other early response genes.

The changes in gene expression for *IL6* and *PGR* were significantly different between Study I and II (Fisher's r-to-z transformation, Table 1). *IL6* expression correlated positively between the two samples within study I but not in study II. This was due to the difference between the D and S samples varying substantially between tumours: there were large increases in IL6 expression in a minority of samples while others remain largely unaffected (Figure 2).

PGR expression was positively correlated between the paired samples in both studies. There was a significant tendency to an increase in study I (expression levels higher in timepoint B than A) and a decrease in study II (expression levels lower in timepoint S than D) that resulted in a marginally significant (p=0.024) heterogeneity between the studies.

Discussion

Multiple issues relating to intra-tumoural heterogeneity are at the forefront of contemporary molecular pathology. One concerns the degree to which a single core-cut biopsy can represent a biomarker's expression across the tumour. We assessed this using a genome-wide approach. We also determined whether two common clinical practices around the time of surgery significantly affected the expression of particular genes or activation of certain pathways. Systematic changes resulting from either process would be relevant to any studies of excised breast cancer, since virtually all excisions occur after diagnostic core-cut and many will involve x-ray of the tumour before its fixation/stabilisation. Data from other studies may differ due to differences between the analytical platforms used.

The variability in whole genome expression data between tissue samples taken peri-surgically has been studied in only small tumour sets (greatest number 13, discussed

below)[4-7]. Pure study of intra-tumoural heterogeneity is best conducted by taking multiple samples from a tumour at the same time. However, the systematic changes occurring in our studies were very modest and will have had little to no perceptible impact on the overall correlations observed. The range of correlations was similar across both studies and overall provided data on 79 tumours. The poorest of the correlations was 0.86 with the large majority being above 0.95 and several being >0.99. Thus gene expression overall shows only modest variability across tumours.

Most investigators are more interested in the variation in expression across the tumour for their gene or genes of interest. Our on-line data [10] will allow them to evaluate that. For illustration we chose 18 genes frequently studied in breast cancer. In general the correlation of the individual genes between the samples was higher for those genes with wide ranges, e.g. *TFF1* (6-log2 range) and *ERBB2* (5-log2 range) than those with narrow ranges, e.g. *SNAI2* (1.5-log2 range) and *MKI67* (<1.0-log2 range). The correlations between individual genes were all worse than those for the genome-wide analyses where there was an approximately 8-log2 range of expression.

We have previously reported that the 60-minute delay in fixation in Study I had no significant impact on immunohistochemical expression of ER, PgR, Ki67, HER2, pAKT or pERK1/2 [2]. Similarly, no genes were found to differ at an FDR<0.05. However, several genes related to stress (e.g. *DUSP1*) and/or known as early response genes (eg *RGS1*, *RGS2*, and *FOSB*) were among those most highly ranked according to change. In Study II, where the larger number of samples provided greater statistical power, the same genes (e.g. *RGS1, FOSB* and *DUSP1*) or similar genes (e.g. *FOS*) ranked in the top 10 genes with changed expression. This suggests that the changes in these early response and stress pathways were true findings in both studies. It is important to note for Study II that no record was made in POETIC of whether excised tumours were subject to x-ray before taking of RNAlater-stored core-cuts. At the Royal Marsden all impalpable tumours and most tumours resected via wide-excision (totalling about 50% of operations) are x-rayed. We have informally determined that similar approaches are in place across the UK. Some of the similarities in the genes changing between the studies may therefore

have been due to a proportion of the tumours in Study II being subjected to x-ray before stabilisation. It should be noted however that while the similarities in the gene changes between the two studies are consistent with delays due to X-ray being responsible in study II there are multiple other factors that occur around surgery that could also contribute. These include the time taken for a sample to reach histopathology where some centres may have taken cores for the POETIC study and delays due to sentinel node biopsy which may have occurred prior to the core being taken. Nonetheless the changes observed in Study II are likely to represent those that occur between diagnostic and surgical samples in common practise and will affect the measurement/study of early response genes in excised tumours.

Two smaller studies have assessed the impact of delay to fixation on global gene expression [4, 5]. In the Borgan study, changes in *FOSB* and *JUND,* while perceptible after 60 minutes, were much greater after 3 hours. The correlation of these changes with time since tumour removal make it likely that they are due to stress of tissue cutting and/or its exposure changed oxygen tension as opposed to the impact of other procedures around surgery such as anaesthesia. The pathway and network analyses undertaken with Study I revealed changes in oxidative phosphorylation and mitochondrial dysfunction. This is also consistent with the exposure of the core-cuts to changed oxygen tension or ischemia. The correlation of mitochondrial dysfunction also correlated quantitatively with time between core-cut taking and fixation supports this change being causatively associated.

Despite the lack of change in the pre-specified immune signatures *IL6* expression increased in both studies and was among the genes identified by Jeselsohn in a similar but smaller study. The change in *IL6* levels in Study II was sufficiently heterogeneous between tumours to nullify the highly significant correlation between the A and B samples in Study I, suggesting that the *IL6* changes were more related to the effects of the initial biopsy than to the short delays around surgery. IL6 is a pleiotropic cytokine secreted by T-cells and macrophages in both systemic and localised immune activation. Its role in breast cancer has been reviewed by Dethlefsen and colleagues [14]. Changes

in *IGFBP2* and particularly *MYC* in Study II also confirmed those seen in the Jeselsohn study, but there was little support for the other 10 genes identified as significant in that study. Like *IL6* these two genes are widely studied in breast cancer. Interpretation of data on them must take account of the effects of diagnostic biopsies.

Some smaller genome–wide analyses between paired biopsies either side of surgery have been reported. Riis et al [7] studied 13 patients with the time between diagnostic and surgical samples ranging between 2 and 8 weeks. As in the current study genes related to early response, including *FOSB* and to oxidative stress including *DUSP1* were differentially expressed between the 2 samples. Similar increases in early response genes including *FOS* were also reported in 16 patients in which fine needle aspirates were taken presurgically and immediately after tumour excision but the time between samples was not stated [8]. Neither of these small studies, identified *IL6, IGFB2* or *MYC* as a changing gene but may have been due to their low statistical power.

There were no systematic differences in categorisation of the tumours into the intrinsic subgroups in either study but discordance was noted between the luminal A versus B subtypes, even after quality control of the RNA and removing technical platform bias with normalization and standardization of expression profiles. In Study II, 15 to 20% of tumours considered luminal A on one core-cut were typed as luminal B or normal-like on the other. Allocation of subtypes is made according to the highest correlation coefficient with the archetypical centroid for each subtype irrespective of the proximity of the correlations to the subtypes although an early report [15] described 43/115 (37%) of tumours as having a low correlation to any of the subtypes. Not surprisingly, we found that subtype discordances were largely associated with small differences between correlations with luminal A and luminal B centroids. The level of discordance in subtyping is important to appreciate given the prominence of intrinsic subtyping in clinical studies of breast cancer and its use for determining whether to allocate chemotherapy [16].

Conclusions

These studies of both random and systematic variability of global gene expression in the context of presurgical study of breast cancer have revealed modest differences in most genes/pathways but confirmed substantial changes in the expression of early response genes that appear to be due to ischemia after surgery and in *IL6, IGFB2* and *MYC* that appear to be responses to initial core-cut biopsy. The data are relevant to all studies of breast cancer since excised tumours almost always have been preceded by core-cut. We provide a reference source [10] for others to assess the potential impact variability in the study of their own genes of interest.

Abbreviations

AGPAT2: 1-Acylglycerol-3-Phosphate O-Acyltransferase 2

ATF3: Activating Transcription factor 3

ATP5C1: ATP Synthase, H+ Transporting, Mitochondrial F1 Complex, Gamma Polypeptide 1

AURKA: Aurora Kinase A

BAG1: BCL2- Associated Athanogene

BHLHB2: Classic B Basic Helix-Loop-Helix Protein 2

DUSP1: Dual Specificity Phosphatase 1

ER: Estrogen Receptor

ERBB2, HER2: Erb-B2 Receptor Tyrosine Kinase 2

FDR: False Discovery Rate

FOS: FBJ Murine Osteosarcoma Viral Oncogene Homolog

FOSB: FBJ Murine Osteosarcoma Viral Oncogene Homolog B

FOXA1: Forkhead Box A1

GADD45B: Growth Arrest And DNA-Damage-Inducible, Beta

GPR160: G Protein-Coupled Receptor 160

HBA2: Hemoglobin, Alpha 2

HBB: Hemoglobin, Beta

HER2-E: Erb-B2 Receptor Tyrosine Kinase 2 Enriched

IL6: Interleukin 6

IGFBP2: Insulin-Like Growth Factor Binding Protein 2

IPA: Ingenuity Pathway Analysis

JUN: Jun Proto-Oncogene

JUND: Jun D Proto-Oncogene

LumA: Luminal A

LumB: Luminal B

MAPT: Microtubule-Associated Protein Tau

MKI67: Marker of Proliferation Ki67

MYC: V-Myc Avian Myelocytomatosis Viral Oncogene Homolog

pAKT: Phospho V-Akt Murine Thymoma Viral Oncogene Homolog 1

pERK1/2: Phospho Extracellular Signal-Regulated Kinase 1/2

PGR: Progesterone Receptor

POETIC: The PeriOperative Endocrine Therapy - Individualising Care

RGS1: Regulator of G-Protein Signaling 1

RGS2: Regulator of G-Protein Signaling 2

SCD: Steroyl-CoA Desaturase

SIK1: Sal-Inducible Kinase 1

SNAI2: Snail Family Zinc Finger 2

SNORD3C: Small Nucleolar RNA, C/D Box 3C

SNORD3D: Small Nucleolar RNA, C/D Box 3D

TFF1: Trefoil Factor 1

TOB1: Transducer of ERBB2, 1

TOP2A: Topoisomerase (DNA) II Alpha

Competing interests

MCU Cheang and J Parker are listed as co-inventor for the PAM50 gene expression classifier patent.

Other authors declare that they have no competing interests.

Authors contributions

ELK extracted RNA from study II, analysed the data and drafted the manuscript. QG analysed the data and drafted the manuscript. MC contributed to the statistical design and interpretation of data, analysis of the intrinsic subtypes and drafting the manuscript. JP did the intrinsic subtype classifier and drafted the manuscript. LAM contributed to the interpretation of data, review and revision of the manuscript. IP assembled samples and extracted RNA from study I. MH, LZ, SD and MA sectioned and reviewed the histopathology of the samples. AD was immunohistochemistry coordinator and allowed data acquisition byreviewing the histopathology. JM provided data and composed table S1. FM contributed to study design and obtained the samples for study I. AS, CH and JR contributed to study conception and provided patient recruitment. IS, JB and MD were involved in conception and design, and drafting of the manuscript. All authors revised and approved the final manuscript.

Acknowledgements

This study was funded in part by Mary-Jean Mitchell Green Foundation, Breast Cancer Now Research Centre. We acknowledge NHS funding to the NIHR Biomedical Research Centre at the Royal Marsden Hospital. The POETIC trial (C1491/A8671/CRUK/07/015, C1491/A15955, C406/A8962), from which samples were obtained for this study, was supported by Cancer Research UK as is ICR-CTSU through its core programme grant.

The study sponsors had no involvement in the design of this study, the literature review, data interpretation, writing of the manuscript or the decision to submit it for publication.

References

1. Dowsett M, Smith I, Robertson J, Robison L, Pinhel I, Johnson L, et al: **Endocrine therapy, new biologicals, and new study designs for presurgical studies**

in breast cancer. *Journal of the National Cancer Institute Monographs* 2011, **2011**(43):120-123.

2. Pinhel IF, MacNeill FA, Hills MJ, Salter J, Detre S, A'Hern R, et al: **Extreme loss of immunoreactive p-Akt and p-Erk1/2 during routine fixation of primary breast cancer**. *Breast Cancer Res* 2010, **12**(5).
3. De Cecco L, Musella V, Veneroni S, Cappelletti V, Bongarzone I, Callari M, et al: **Impact of biospecimens handling on biomarker research in breast cancer**. *BMC Cancer* 2009, **9**:409.
4. Borgan E, Navon R, Vollan HK, Schlichting E, Sauer T, Yakhini Z, et al: **Ischemia caused by time to freezing induces systematic microRNA and mRNA responses in cancer tissue**. *Mol Oncol* 2011, **5**(6):564-576.
5. Aktas B, Sun H, Yao H, Shi W, Hubbard R, Zhang Y, et al: **Global gene expression changes induced by prolonged cold ischemic stress and preservation method of breast cancer tissue**. *Mol Oncol* 2014, **8**(3):717-727.
6. Jeselsohn RM, Werner L, Regan MM, Fatima A, Gilmore L, Collins LC, et al: **Digital quantification of gene expression in sequential breast cancer biopsies reveals activation of an immune response**. *PLoS One* 2013, **8**(5):e64225.
7. Riis ML, Luders T, Markert EK, Haakensen VD, Nesbakken AJ, Kristensen VN, et al: **Molecular profiles of pre- and postoperative breast cancer tumours reveal differentially expressed genes**. *ISRN Oncol* 2012, **2012**:450267.
8. Wong V, Wang DY, Warren K, Kulkarni S, Boerner S, Done SJ, et al: **The effects of timing of fine needle aspiration biopsies on gene expression profiles in breast cancers**. *BMC Cancer* 2008, **8**:277.
10. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73237
11. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, et al: **Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival**. *Proc Natl Acad Sci U S A* 2005, **102**(10):3738-3743.

12. Dunbier AK, Ghazoui Z, Anderson H, Salter J, Nerurkar A, Osin P, et al: **Molecular profiling of aromatase inhibitor-treated postmenopausal breast tumors identifies immune-related correlates of resistance**. *Clin Cancer Res* 2013, **19**(10):2775-2786.

13. Fisher RA: **Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population**. *Biometrika* 1915, **10**(4):507-521.

14. Dethlefsen C, Hojfeldt G, Hojman P: **The role of intratumoral and systemic IL-6 in breast cancer**. *Breast Cancer Res Treat* 2013, **138**(3):657-664.

15. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8418-8423.

16. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al: **Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013.** *Ann Oncol* 2013, **24**(9):2206-2223.

Figure legends

Figure 1. Hierarchical clustering with Euclidean distance and average linkage, based on (A) Study I. Clustering of 24,395 probes and 23 pairs of samples; B) Study II. Clustering of 32,332 probes and 56 pairs of samples. In brief, probes and samples were grouped based on similarities calculated using the Euclidean distance method and average linkage (Additional file 1. Supplementary information). Sample dendrogram bars were coloured according to PAM50 intrinsic subtypes and Pairing of samples respectively. PAM50 color: green = Normal; dark blue = LumA; light blue = LumB; purple = Her2-enriched; red = Basal; grey = Paired together: light green = Unpaired first sample; dark green = Unpaired second sample.

Figure 2. Line Diagram of the paired IL6 expression levels in Study I and Study II. Study I IL6 expression levels of samples A and B and Study II IL6 expression levels at diagnosis (D) and surgery (S). Marked in red are samples with >50% increase in expression.

Tables

Table 1. Correlation of paired expression levels in 5 genes reported in breast cancer (complete list of 18 genes in Table S2) and 9 genes identified by Jeselsohn

Table 2. Top 8 genes significantly different in paired samples of Study I and Study II

Additional files

Additional file 1. Supplementary Information. Additional description of the materials and methods (.doc).

Additional file 2. Figure S1. Paired correlations in Study I. Correlation of detectable probes by Pearson correlation in 23 pairs of samples (.pdf).

Additional file 2. Figure S2. Examples of paired correlations in Study I. Correlation of detectable probes by Pearson correlation: the 3 samples with the highest correlations, median correlation and the lowest correlations **(.pdf)**.

Additional file 2. Figure S3. Correlation of 18 genes in Study I. Pearson correlation of 18 genes commonly studied in breast cancer in 23 pairs of samples (.pdf).

Additional file 2. Figure S4. Scatterplots of numeric differences between correlation coefficients to average gene expression profiles of Intrinsic subtypes for each tumor in Study I (S4A and B) and Study II (S4C and D). Difference between Luminal A and Luminal B centroids (A and C), and Luminal B and HER2-Enriched centroids (C and D). Open circle: concordant subtype assignments between the two time points. Triangle: discordant subtype assignments between the two time points (.pdf).

Additional file 2. Figure S5. Paired correlations in Study II. Correlation of detectable probes by Pearson correlation in 56 pairs of samples (.pdf).

Additional file 2. Figure S6. Correlations of 18 genes in Study II. Pearson correlation of 18 genes commonly studied in breast cancer in 56 pairs of samples (.pdf).

Additional file 3. Table S1. Demographics for Study I and Study II

Additional file 3. Table S2. Correlation of paired expression levels in 13 genes reported in breast cancer (complementing Table 1).

Additional file 3. Table S3. A) Canonical pathways and B) Top networks identified in Study I.

Additional file 3. Table S4. Genes correlated with time elapsed in Study I.

Additional file 3. Table S5. Top pathways identified from 116 genes correlated with time elapsed.

Additional file 3. Table S6. Top networks identified from 116 genes correlated with time elapsed.

Additional file 3. Table S7. Intrinsic subtype concordance between pairs.

Additional file 3. Table S8. Top pathways identified in Study II.

Additional file 3. Table S9. Top networks identified in Study II.

Additional file 3. Table S10. Top 20 genes identified in Study I and their p-value in Study II.

Table 1. Correlation of paired expression levels in 5 genes reported in breast cancer and 9 genes identified by Jeselsohn.

			STUDY I					STUDY II					STUDY I vs. STUDY II	
		Gene symbol	**R**	**P value**	**Geometric Mean of B/A**	**95% CI**		**R**	**P value**	**Geometric Mean of S/D**	**95% CI**		**Z-value**	**P-value (2 tail)**
		BAG1	0.713	0.0001	0.971	**0.946-0.996**		0.734	<0.0001	1.043	0.984-1.106		-0.17	0.865
		MKi67	0.354	0.0978	1.009	0.962-1.058		0.522	<0.0001	0.977	0.930-1.027		-0.8	0.4237
		MAPT	0.847	<0.0001	0.806	**0.692-0.938**		0.811	<0.0001	1.108	0.965-1.273		0.44	0.6599
		PGR	0.522	0.0106	1.093	0.946-1.263		0.824	<0.0001	0.978	0.894-1.070		**-2.25**	**0.0244**
		SNAI2	0.430	0.0408	0.897	0.790-1.018		0.481	0.0002	0.940	0.838-1.054		-0.25	0.8026
Genes that significantly changed in Jeselsohn et al (2013)	(a) immune related	IGFBP2	0.583	0.0035	1.051	0.862-1.282		0.784	<0.0001	1.136	**1.031-1.251**		-1.48	0.1389
		IL6	0.712	0.0001	1.108	**1.003-1.223**		0.194	0.1525	1.167	**1.079-1.262**		**2.65**	**0.008**
		CD68	0.412	0.0509	1.065	0.889-1.272		0.464	0.0003	1.099	0.985-1.226		-0.25	0.8026
		CD14	0.553	0.0062	1.047	0.905-1.211		0.355	0.0074	1.017	0.901-1.148		0.96	0.3371
		CD52	0.755	< 0.0001	1.085	0.923-1.276		0.436	0.0008	1.038	0.876-1.230		1.97	0.0488
		CD44	0.458	0.0278	0.927	0.788-1.091		0.816	<0.0001	0.952	0.890-1.019		-2.48	0.0131
		PPARG	0.315	0.1438	0.806	0.608-1.068		0.343	0.0096	0.993	0.870-1.132		-0.12	0.9045
		ADM	0.476	0.0217	0.931	0.720-1.204		0.544	<0.0001	1.122	0.964-1.306		-0.35	0.7263
		VEGFA	0.653	0.0007	1.043	0.967-1.124		0.647	<0.0001	0.991	0.930-1.055		0.04	0.9681
	(b) non-immune related	CENPF	0.781	< 0.0001	1.039	0.913-1.183		0.729	<0.0001	1.062	0.959-1.176		0.46	0.6455
		MYC	0.509	0.0132	1.076	0.897-1.292		0.65	<0.0001	1.439	**1.241-1.668**		-0.82	0.4122
		CCNB1	0.413	0.0501	0.976	0.883-1.078		0.469	0.0003	1.010	0.919-1.107		-0.27	0.7872
		MAP1LC3B	0.598	0.0026	0.957	0.882-1.038		0.809	<0.0001	0.971	0.933-1.010		-1.65	0.099
		SNAI1	ND	ND	ND	ND		ND	ND	ND	ND		ND	ND

ND=non-Detected

Table 2. Top 8 genes significantly different in paired samples of Study I and Study II

STUDY I						STUDY II				
Accession	**Symbol**	**Parametric p-value**	**FDR**	**FC**		**Accession**	**Symbol**	**Parametric p-value**	**FDR**	**FC**
NM_006732	FOSB	0.0014	0.138	2.08		NM_005252	FOS	< 1e-07	< 1e-07	4.00
NM_004417	DUSP1	0.0003	0.133	1.72		NM_002922	RGS1	< 1e-07	< 1e-07	3.23
NM_002923	RGS2	0.0003	0.133	1.59		NM_004417	DUSP1	< 1e-07	< 1e-07	3.13
NM_003407	ZFP36	0.0005	0.133	1.54		NM_000517	HBA2	< 1e-05	0.003	-2.90
NM_033027	AXUD1	0.0001	0.087	1.49		NM_000518	HBB	< 1e-05	0.006	-2.83
NM_004566	PFKFB3	0.0030	0.153	-1.48		NM_000517	HBA2	< 1e-05	0.007	-2.64
NM_018955	UBB	0.0037	0.155	-1.46		NM_000558	HBA1	< 1e-04	0.008	-2.39
NM_005063	SCD	0.0003	0.133	-1.45		NM_006732	FOSB	< 1e-06	0.001	2.38

Additional file 1. Supplementary Information

Gene expression analysis and data pre-processing

Total RNA was extracted using miRNeasy (Qiagen, Sussex, UK). RNA quality was checked using an Agilent Bioanalyser (Santa Clara, CA, USA): samples with RNA integrity values of <4 were excluded from further analysis. RNA amplification, labelling and hybridization on HumanHT-12_V3 (study I samples) and HumanHT-12_V4 (study II samples) expression BeadChips (Illumina, San Diego, CA, USA) were performed according to the manufacturer's instructions. Illumina raw data was extracted using GenomeStudio software and was transformed and normalized using variance-stabilizing transformation, robust spline normalization method included in the R package (lumi) (http://www.bioconductor.org). The data was then batch-corrected using the function (ComBat) in the R package (sva). Paired samples were excluded from further analysis if their fraction of detected genes was <30% and identified as outliers by a sample outlier detection function in the lumi package. Probes were filtered out if they were not detected in any of the samples (detection p>0.01). Gene expression data from this study is deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73237) with accession number GSE73237.

Gene Signatures

Entrez Gene ID was used as gene identifier in gene signatures. The HumanHT-12_V4_0_R2_15002873_B annotation file was used to map the EntrezGeneIDs to the corresponding Illumina probe IDs. Gene signature scores were weighted averages as described previously [1].

We evaluated three candidate gene sets: i) metagene wound healing signature [2]; ii) immune response metagene [3] and iii) 13 of the 14 genes identified as changing in the Jeselsohn study [4] (SNAI1 was not detected on the Illumina

platform). We also studied the effects on 18 pre-specified genes that we selected as being particularly relevant to breast cancer from prior studies.

Each tumour sample was classified into one of the five intrinsic subtypes based on the PAM50 classifier [5]. Prior to classification, technical bias between these data and the training data were minimized to ensure accurate calls across heterogeneous platforms. Under the assumption that The Cancer Genome Atlas ER+ cohort and the baseline specimens of the POETIC cohort were similar, gene-wise differences in the mean and variance of these two groups represent technical bias. These differences were removed from the POETIC and study I cohorts prior to PAM50 classification respectively.

Single sample intrinsic subtype prediction was performed by calculating a Spearman rank correlation coefficient between the 50-gene expression values of an individual sample compared to each of the average gene expression (centroid) values for Luminal A, Luminal B, HER2-Enriched, Basal-like, and Normal. The subtype classification for the study sample is assigned to the centroid with the highest correlation.

Data analysis and Statistical Methods

Pearson correlation coefficient was used to assess the association of the: i) detectable probes between the paired samples and ii) pre-selected genes between paired samples' expression levels. Univariate paired or unpaired T-tests together with multivariate permutation tests were used to identify differentially expressed genes between the paired samples. The significantly differentially expressed genes were subjected to Ingenuity Pathway Analysis (IPA). Pathways were considered as significantly altered if $p<0.05$ after using Benjamini-Hochberg Multiple Testing Correction. Wilcoxon matched-pairs signed rank test was used to

evaluate the significance of the percentage increase of expression between pairs. The correlation of the difference in gene expression between biopsies and the length of the time interval between the biopsies was evaluated using Spearman rank correlation. The significance of the difference between 2 correlation coefficients obtained in study I and study II respectively was calculated using the Fisher r-to-z transformation [6] using the online calculator (http://vassarstats.net/rdiff.html). For each sample we calculated the following: (a) numeric difference of the LumA and LumB centroid correlation coefficients (i.e. LumA correlation coefficient minus Lum B correlation coefficient) and (b) numeric difference of their LumB and HER2-enriched correlation coefficients (i.e. LumB correlation coefficient minus HER2-enriched subtype correlation coefficient). Medians of these centroid correlation coefficients were reported and approximate 95% C.I. intervals were calculated using the adjusted bootstrap percentile method where appropriate [7]. No formal statistics comparison of the medians is performed.

GraphPad Prism 6 (Graphpad Software Inc.) was used for some of the statistical analyses in this study.

Hierarchical clustering method

To identify the clusters in gene expression between samples we used the Euclidean distance method and average linkage of sampleRelation function within the “lumi” R-package. The samples were then color annotated according to their PAM50 intrinsic subtypes (green = Normal; dark blue = LumA; light blue = LumB; purple = Her2-enriched; red = Basal) and whether or not the paired samples were clustered together (grey = Paired together: light green = Unpaired first sample; dark green = Unpaired second sample). In the final heatmaps of gene expression for the probes, they were generated based on the same clustering method (i.e. Euclidean distance method and average linkage), but

keeping the order of samples. Function named colorRampPalette within R-package was used to specify the gradient of the colors.

References

1. Gao Q, Patani N, Dunbier AK, Ghazoui Z, Zvelebil M, Martin LA, Dowsett M: **Effect of aromatase inhibition on functional gene modules in estrogen receptor-positive breast cancer and their relationship with antiproliferative response**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2014, **20**(9):2485-2494.
2. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H *et al*: **Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(10):3738-3743.
3. Dunbier AK, Ghazoui Z, Anderson H, Salter J, Nerurkar A, Osin P, A'Hern R, Miller WR, Smith IE, Dowsett M: **Molecular profiling of aromatase inhibitor-treated postmenopausal breast tumors identifies immune-related correlates of resistance**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2013, **19**(10):2775-2786.
4. Jeselsohn RM, Werner L, Regan MM, Fatima A, Gilmore L, Collins LC, Beck AH, Bailey ST, He HH, Buchwalter G *et al*: **Digital quantification of gene expression in sequential breast cancer biopsies reveals activation of an immune response**. *PloS one* 2013, **8**(5):e64225.
5. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised risk predictor of breast cancer based on intrinsic subtypes**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009, **27**(8):1160-1167.

6. Fisher RA: **Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population**. *Biometrika* 1915, **10**(4):507-521.

7. Efron B: **Better Bootstrap Confidence Intervals**. *Journal of the American Statistical Association* 1987, **82**(397):171-185.

Figure S1

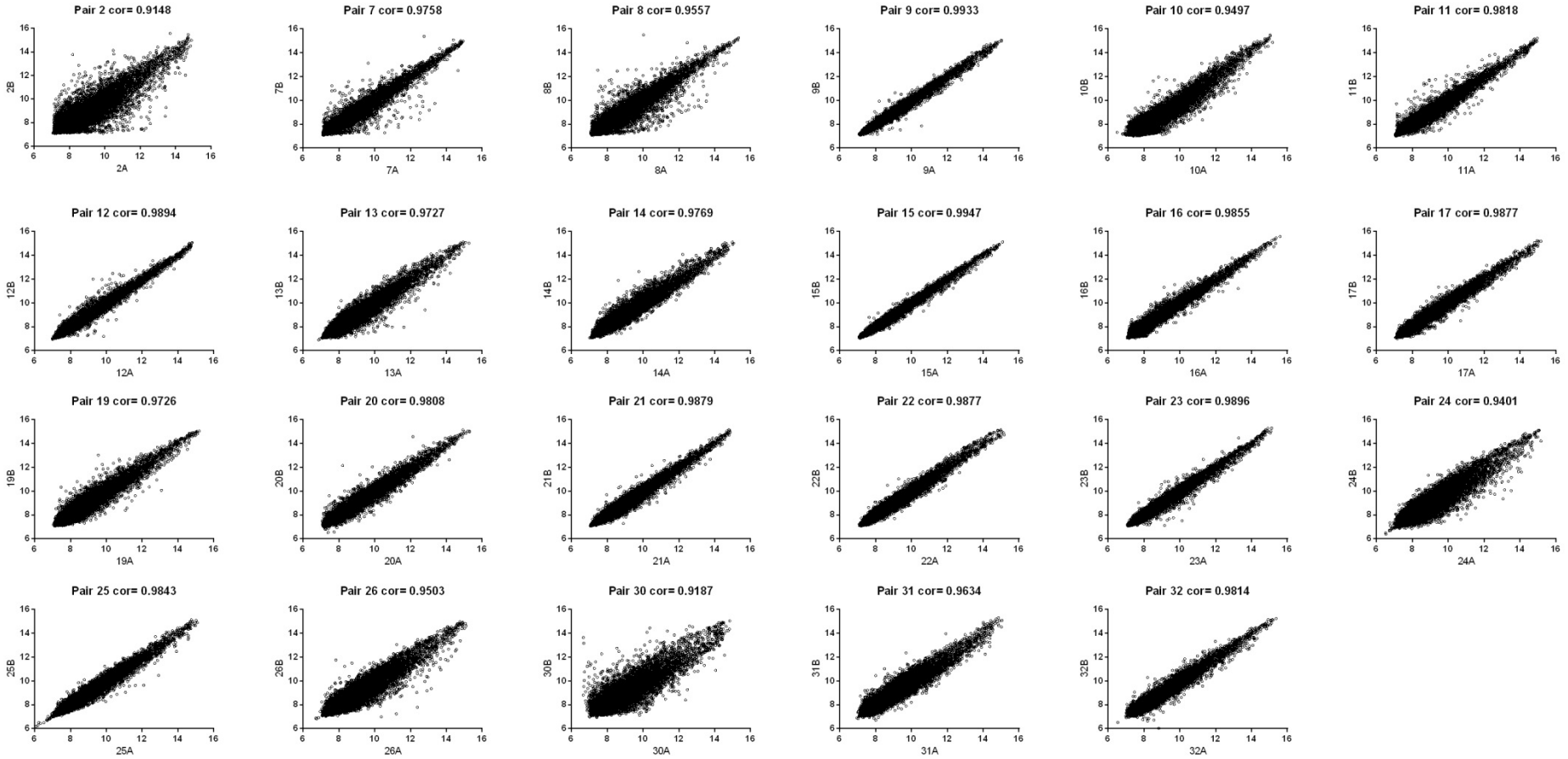

Pair 2 cor= 0.9148
Pair 7 cor= 0.9758
Pair 8 cor= 0.9557
Pair 9 cor= 0.9933
Pair 10 cor= 0.9497
Pair 11 cor= 0.9818
Pair 12 cor= 0.9894
Pair 13 cor= 0.9727
Pair 14 cor= 0.9769
Pair 15 cor= 0.9947
Pair 16 cor= 0.9855
Pair 17 cor= 0.9877
Pair 19 cor= 0.9726
Pair 20 cor= 0.9808
Pair 21 cor= 0.9879
Pair 22 cor= 0.9877
Pair 23 cor= 0.9896
Pair 24 cor= 0.9401
Pair 25 cor= 0.9843
Pair 26 cor= 0.9503
Pair 30 cor= 0.9187
Pair 31 cor= 0.9634
Pair 32 cor= 0.9814

Figure S2

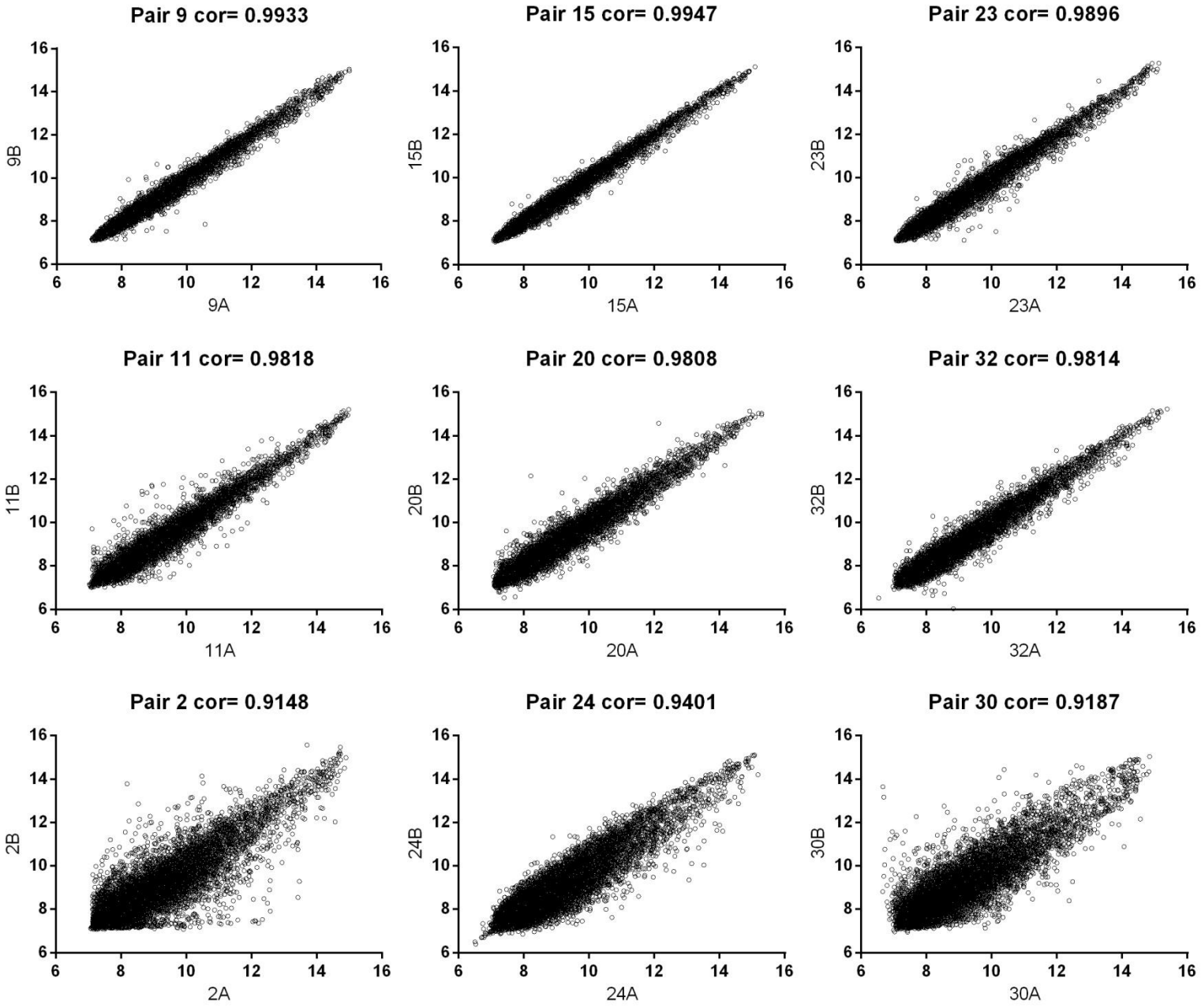
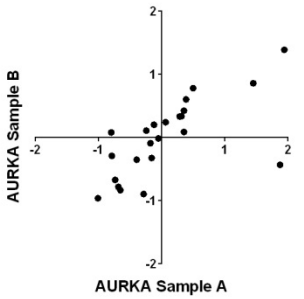
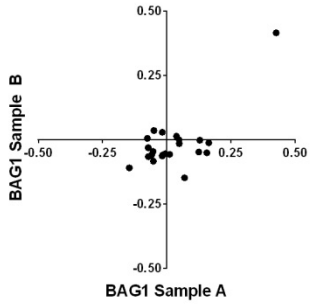
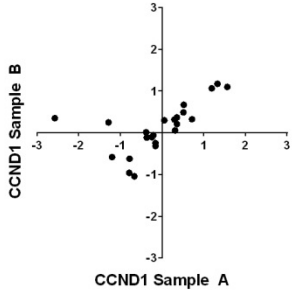

Pair 9 cor= 0.9933
9B
9A
Pair 15 cor= 0.9947
15B
15A
Pair 23 cor= 0.9896
23B
23A
Pair 11 cor= 0.9818
11B
11A
Pair 20 cor= 0.9808
20B
20A
Pair 32 cor= 0.9814
32B
32A
Pair 2 cor= 0.9148
2B
2A
Pair 24 cor= 0.9401
24B
24A
Pair 30 cor= 0.9187
30B
30A
6
8
10
12
14
16

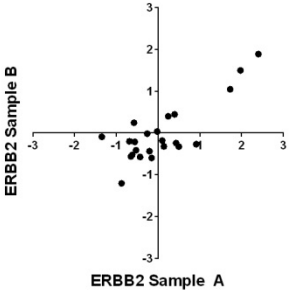
Figure S3

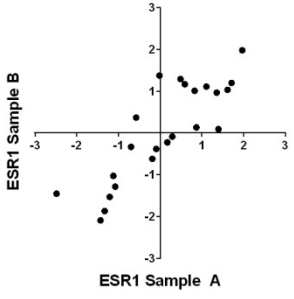
Study I

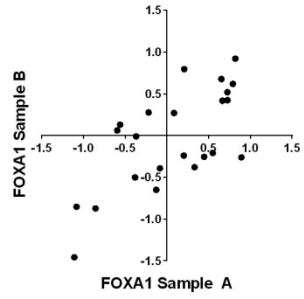


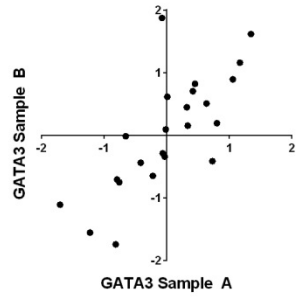


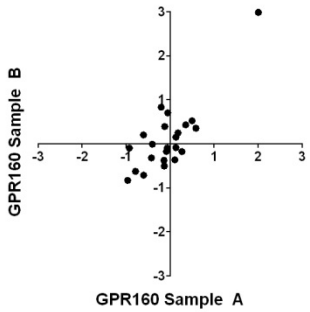


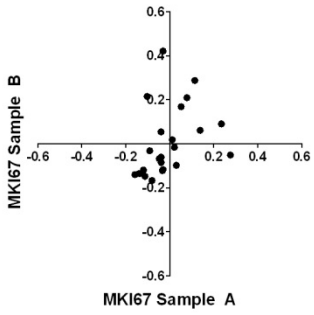


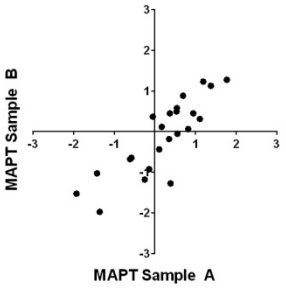


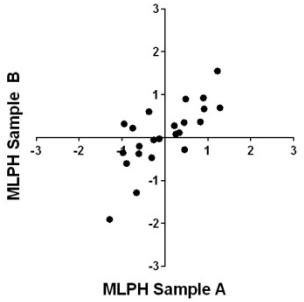


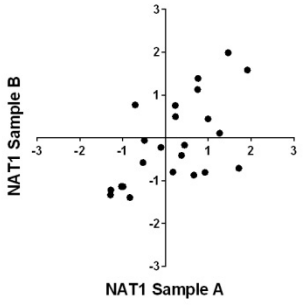


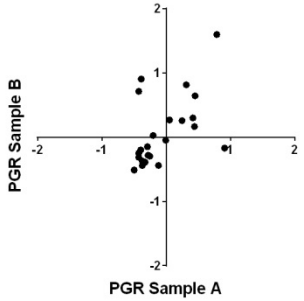


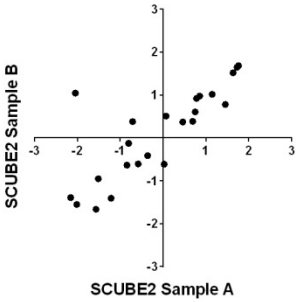


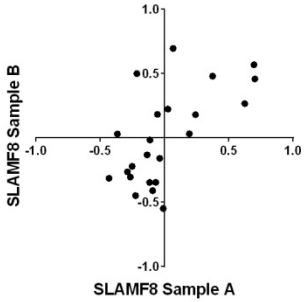


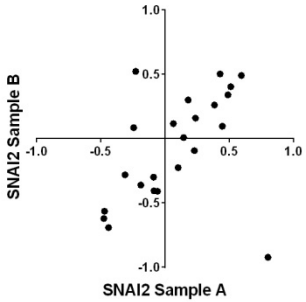


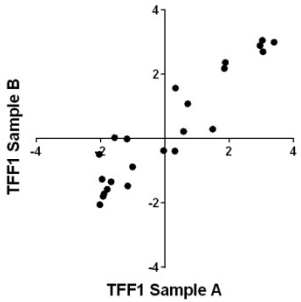












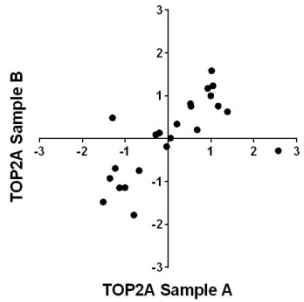
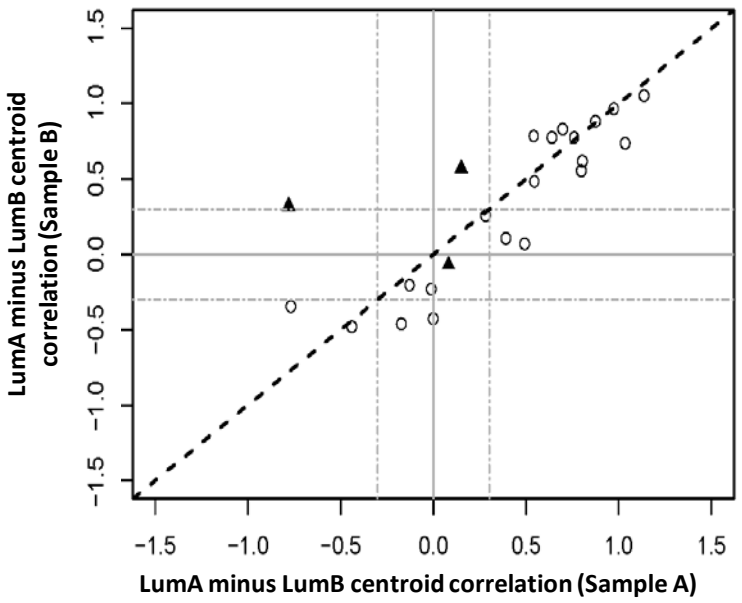
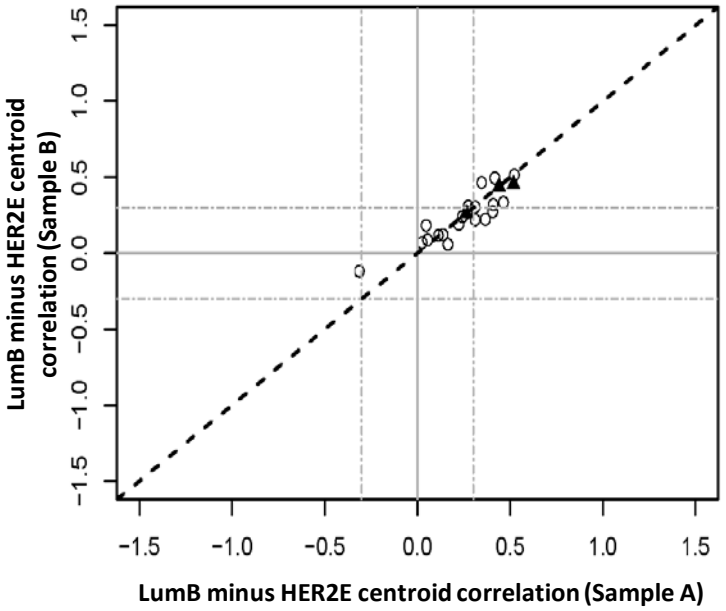


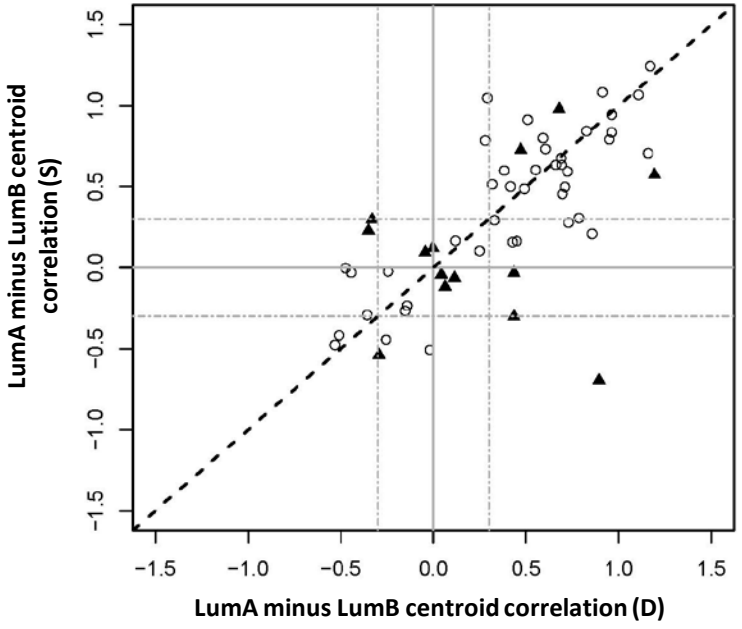
Figure S4

A. Study I





B. Study II



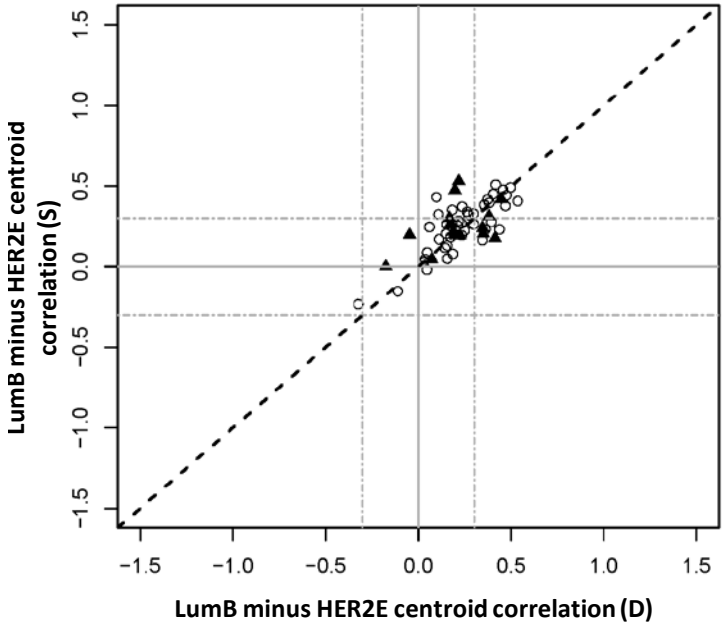


Figure S5

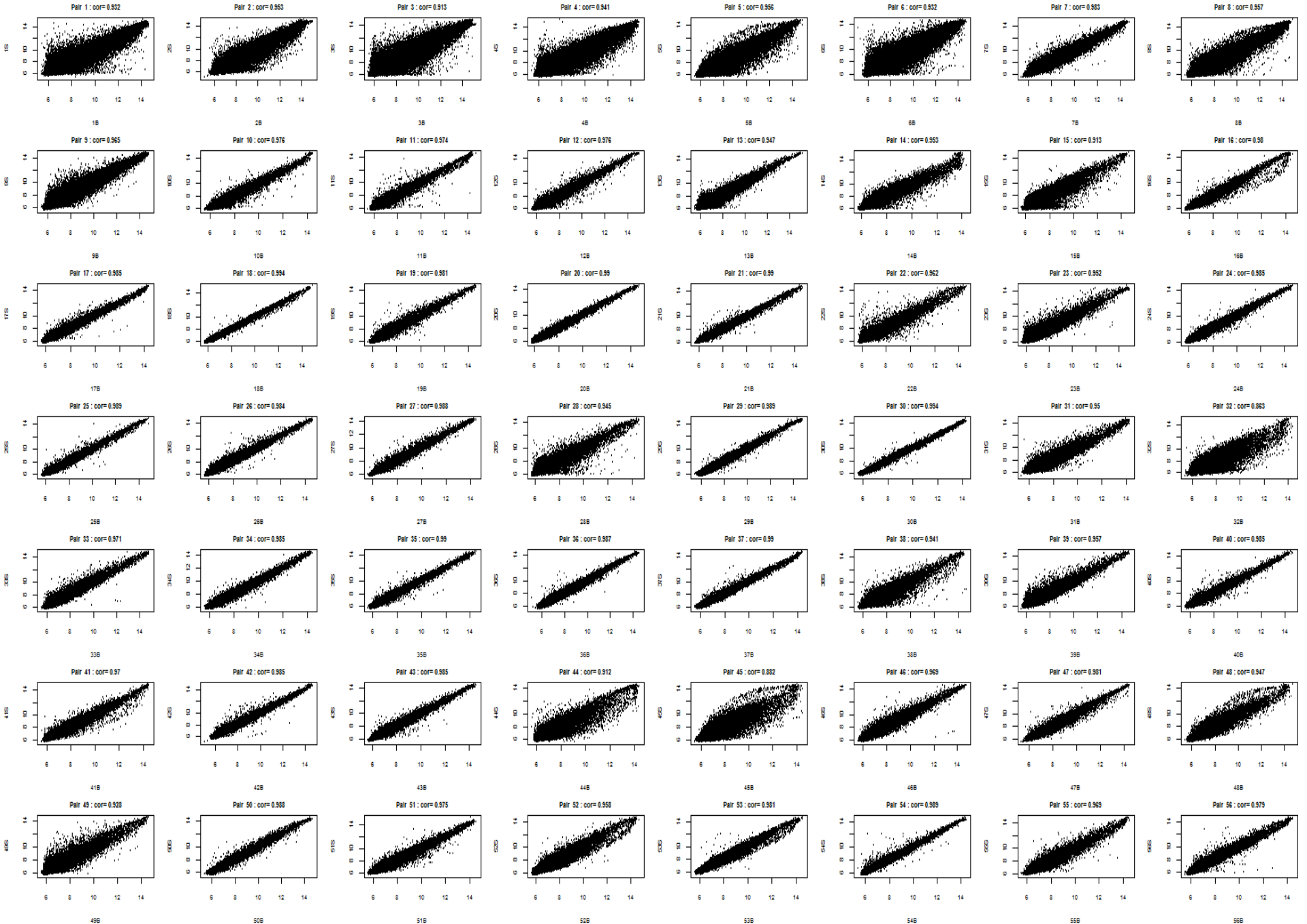

Pair 1 : cor= 0.932
Pair 2 : cor= 0.953
Pair 3 : cor= 0.913
Pair 4 : cor= 0.941
Pair 5 : cor= 0.956
Pair 6 : cor= 0.932
Pair 7 : cor= 0.983
Pair 8 : cor= 0.957
Pair 9 : cor= 0.965
Pair 10 : cor= 0.976
Pair 11 : cor= 0.974
Pair 12 : cor= 0.976
Pair 13 : cor= 0.947
Pair 14 : cor= 0.953
Pair 15 : cor= 0.913
Pair 16 : cor= 0.98
Pair 17 : cor= 0.985
Pair 18 : cor= 0.994
Pair 19 : cor= 0.981
Pair 20 : cor= 0.99
Pair 21 : cor= 0.99
Pair 22 : cor= 0.962
Pair 23 : cor= 0.952
Pair 24 : cor= 0.985
Pair 25 : cor= 0.989
Pair 26 : cor= 0.984
Pair 27 : cor= 0.988
Pair 28 : cor= 0.945
Pair 29 : cor= 0.989
Pair 30 : cor= 0.994
Pair 31 : cor= 0.95
Pair 32 : cor= 0.863
Pair 33 : cor= 0.971
Pair 34 : cor= 0.985
Pair 35 : cor= 0.99
Pair 36 : cor= 0.987
Pair 37 : cor= 0.99
Pair 38 : cor= 0.941
Pair 39 : cor= 0.957
Pair 40 : cor= 0.985
Pair 41 : cor= 0.97
Pair 42 : cor= 0.985
Pair 43 : cor= 0.985
Pair 44 : cor= 0.912
Pair 45 : cor= 0.882
Pair 46 : cor= 0.969
Pair 47 : cor= 0.981
Pair 48 : cor= 0.947
Pair 49 : cor= 0.928
Pair 50 : cor= 0.988
Pair 51 : cor= 0.975
Pair 52 : cor= 0.958
Pair 53 : cor= 0.981
Pair 54 : cor= 0.989
Pair 55 : cor= 0.969
Pair 56 : cor= 0.979

Figure S6

Study II

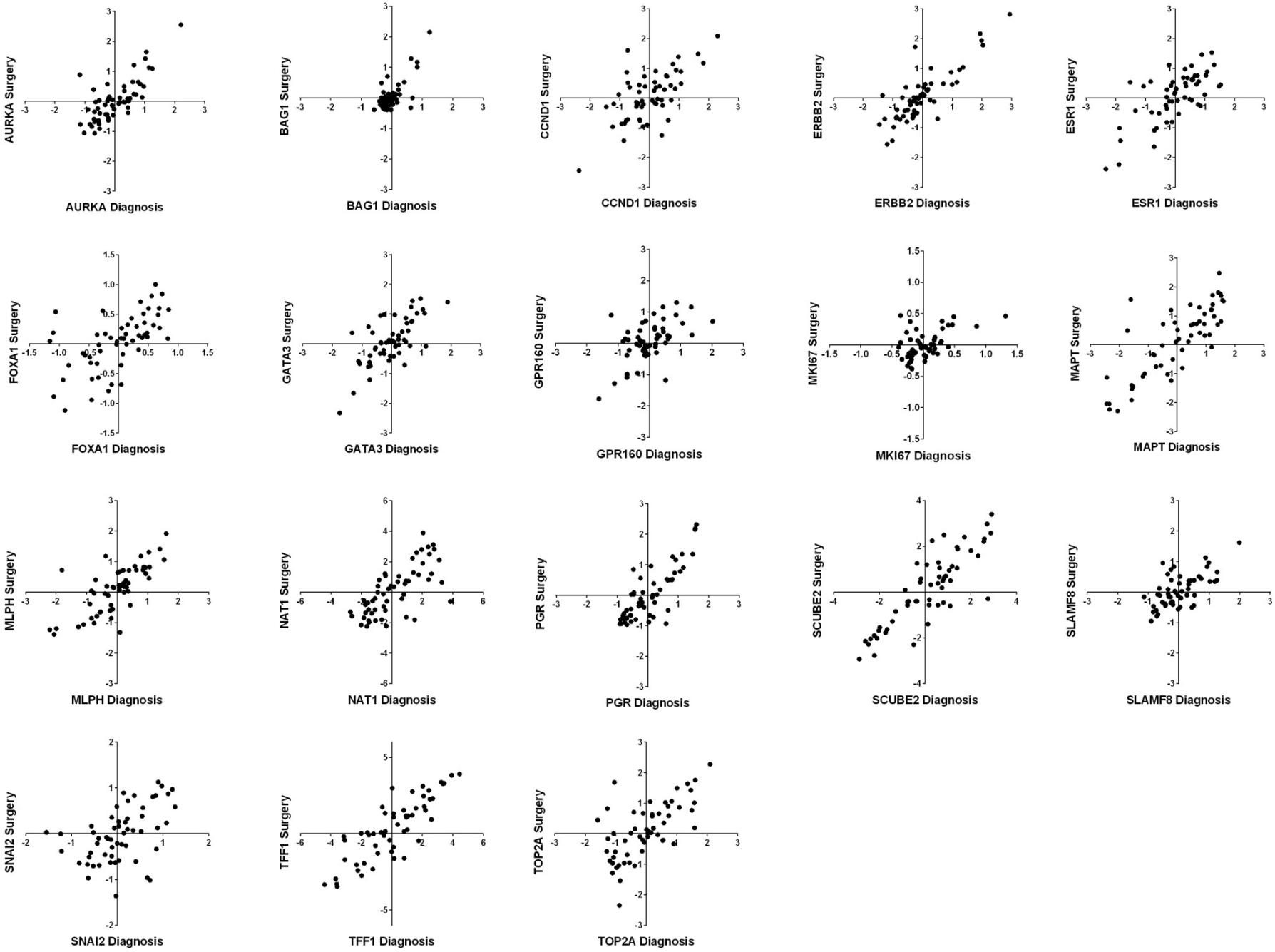


Table S1. Demographics for Study I and Study II

	Study I		Study II	
	n	%	**n**	**%**
Age at randomisation (years)				
50-59	12	52.2	4	7.1
60-69	5	45.5	16	28.6
70-79	2	3.3	16	28.6
≥80	4	6.8	20	35.7
Age at randomisation - Median (IQR)	55	(47-70)	76.6	(67.4 - 81.7)
Tumour grade				
G1	6	26.1	8	14.3
G2	8	34.8	30	53.6
G3	8	34.8	8	14.3
Not known	1	4.3	10	17.9
Tumour size (cm)				
≤2	5	21.7	14	25
>2 & ≤5	14	60.9	41	73.2
>5	4	17.4	1	1.8
Nodal status				
Negative	14	60.9	33	58.9
Positive	8	34.8	23	41.1
Not known	1	4.3		
Histological type				
Ductal	18	78.3	43	76.8
Lobular	3	13.0	9	16.1
DCIS	2	8.7		
Mucinous			3	5.4
Mixed ductal and lobular			1	1.8
ER status				
Positive	21	91.3	56	100
Negative	2	8.7		
PgR status				
Positive	19	82.6	44	78.6
Negative	4	17.4	6	10.7
Not known			6	10.7
HER2 status				
Negative	20	87.0	46	82.1
Positive	3	13.0	8	14.3
Not known			2	3.6
Ki67 (%) at baseline - Median (IQR)*	14.2	6.9-17.1	19	11.9 - 33.6

Nodal status and HER status are recorded post-surgery, all other characteristics recorded pre-surgery. Tumour size is measured either by ultrasound or clinical examination

*Baseline Ki67 data unavailable for 5/56 patients

Table S2. Correlation of paired expression levels in 18 genes reported in breast cancer and

		STUDY I					
		Gene symbol	**R**	**P value**	**Geometric Mean of B/A**	**95% CI**	**R**
Genes selected as commonly studied in breast cancer		AURKA	0.677	0.0004	0.951	0.796-1.137	0.759
		BAG1	0.713	0.0001	0.971	**0.946-0.996**	0.734
		CCND1	0.621	0.0016	1.133	0.912-1.408	0.645
		ERBB2	0.811	<0.0001	0.926	0.786-1.091	0.844
		ESR1	0.847	<0.0001	0.958	0.787-1.165	0.715
		FOXA1	0.686	0.0003	0.922	0.796-1.067	0.597
		GATA3	0.756	<0.0001	1.018	0.847-1.223	0.704
		GPR160	0.805	<0.0001	1.118	0.975-1.282	0.554
		MKi67	0.354	0.0978	1.009	0.962-1.058	0.522
		MAPT	0.847	<0.0001	0.806	**0.692-0.938**	0.811
		MLPH	0.741	<0.0001	1.06	0.901-1.246	0.741
		NAT1	0.604	0.0023	0.813	0.626-1.056	0.717
		PGR	0.522	0.0106	1.093	0.946-1.263	0.824
		SCUBE2	0.806	<0.0001	1.158	0.923-1.453	0.857
		SLAMF8	0.621	0.0016	0.996	0.909-1.090	0.655
		SNAI2	0.43	0.0408	0.897	0.790-1.018	0.481
		TFF1	0.932	<0.0001	1.148	0.932-1.413	0.842
		TOP2A	0.682	0.0003	0.977	0.766-1.247	0.651
Genes that significantly changed in Jeselsohn et al (2013)	(a) immune related	IGFBP2	0.583	0.0035	1.051	0.862-1.282	0.784
		IL6	0.712	0.0001	1.108	**1.003-1.223**	0.194
		CD68	0.412	0.0509	1.065	0.889-1.272	0.464
		CD14	0.553	0.0062	1.047	0.905-1.211	0.355
		CD52	0.755	< 0.0001	1.085	0.923-1.276	0.436
		CD44	0.458	0.0278	0.927	0.788-1.091	0.816
		PPARG	0.315	0.1438	0.806	0.608-1.068	0.343
		ADM	0.476	0.0217	0.931	0.720-1.204	0.544
		VEGFA	0.653	0.0007	1.043	0.967-1.124	0.647
	(b) non-immune related	CENPF	0.781	< 0.0001	1.039	0.913-1.183	0.729
		MYC	0.509	0.0132	1.076	0.897-1.292	0.65
		CCNB1	0.413	0.0501	0.976	0.883-1.078	0.469
		MAP1LC3B	0.598	0.0026	0.957	0.882-1.038	0.809
		SNAI1	ND	ND	ND	ND	ND

9 identified by Jeselsohn.

STUDY II				STUDY I vs. STUDY II	
P value	**Geometric Mean of S/D**	**95% CI**		**Z-value**	**P-value (2 tail)**
<0.0001	1.01	0.923-1.106		-0.65	0.5157
<0.0001	1.043	0.984-1.106		-0.17	0.865
<0.0001	1.048	0.920-1.194		-0.15	0.8808
<0.0001	1.065	0.973-1.166		-0.4	0.6892
<0.0001	1.027	0.910-1.159		1.33	0.1835
<0.0001	1.037	0.952-1.129		0.58	0.5619
<0.0001	1.083	0.974-1.203		0.43	0.6672
<0.0001	1.034	0.923-1.159		1.86	0.0629
<0.0001	0.977	0.930-1.027		-0.8	0.4237
<0.0001	1.108	0.965-1.273		0.44	0.6599
<0.0001	1.107	0.992-1.235		0	1
<0.0001	0.944	0.750-1.188		-0.77	0.4413
<0.0001	0.978	0.894-1.070		**-2.25**	**0.0244**
<0.0001	0.989	0.846-1.156		-0.63	0.5287
<0.0001	1.027	0.935-1.129		-0.22	0.8259
0.0002	0.94	0.838-1.054		-0.25	0.8026
<0.0001	1.216	0.980-1.509		1.7	0.0891
<0.0001	1.089	0.944-1.255		0.21	0.8337
<0.0001	1.136	**1.031-1.251**		-1.48	0.1389
0.1525	1.167	**1.079-1.262**		**2.65**	**0.008**
0.0003	1.099	0.985-1.226		-0.25	0.8026
0.0074	1.017	0.901-1.148		0.96	0.3371
0.0008	1.038	0.876-1.230		1.97	0.0488
<0.0001	0.952	0.890-1.019		-2.48	0.0131
0.0096	0.993	0.870-1.132		-0.12	0.9045
<0.0001	1.122	0.964-1.306		-0.35	0.7263
<0.0001	0.991	0.930-1.055		0.04	0.9681
<0.0001	1.062	0.959-1.176		0.46	0.6455
<0.0001	1.439	**1.241-1.668**		-0.82	0.4122
0.0003	1.01	0.919-1.107		-0.27	0.7872
<0.0001	0.971	0.933-1.010		-1.65	0.099
ND	ND	ND		ND	ND

Table S3. A) Canonical pathways and B) Top networks

Table S3A

Canonical Pathways	B-H p-value
Oxidative Phosphorylation	8.9125E-05
Mitochondrial Dysfunction	0.00083176
CDK5 Signaling	0.01479108
Oleate Biosynthesis II (Animals)	0.01659587
Protein Ubiquitination Pathway	0.03801894
Aldosterone Signaling in Epithelial Cells	0.03801894

Table S3B

ID	Score
1	39
2	31
3	28
4	21
5	8
6	2

identified in Study I.

Molecules
CYB5A,ATP5O,ATP5F1,ATP5J,ATP5C1,COX5B
CYB5A,ATP5O,ATP5F1,ATP5J,ATP5C1,COX5B
GNAS,LAMB1,FOSB,PPP1R1B
CYB5A,SCD
UBB,PSMA3,PSMC6,HSP90AA1,HSPE1
DUSP1,SGK1,HSP90AA1,HSPE1

Focus Molecules	Top Diseases and Functions
18	DNA Replication, Recombination, and Repair, Nucleic Acid Metabolism, Small Molecule Biochemistry
15	DNA Replication, Recombination, and Repair, Energy Production, Nucleic Acid Metabolism
14	Developmental Disorder, Hereditary Disorder, Metabolic Disease
11	Cell Death and Survival, Embryonic Development, Cellular Movement
5	Cell Signaling, Molecular Transport, Nucleic Acid Metabolism
1	Cancer, Organismal Injury and Abnormalities, Reproductive System Disease

Molecules in Network
20s proteasome,26s Proteasome,ADCY,BST2,Calcineurin protein(s),Cg,Creb,DUSP1,ERK1/2,FOSB,FSH,GNAS,hemoglobin,HSPE1,Insulin,KLF13,Lh,MAP2K1/2,NR4A2,Pde,PDE5A,PDGF BB,PDXK,PFKFB3,phosphatase,Pkg,PPP1R1B,PRDX2,Pro-inflammatory Cytokine,PSMC6,PTPLB,RETSAT,RGS2,ZAK,ZFP36
adenosine-tetraphosphatase,AGPAT2,ARRB1,ARRDC1,ATP synthase,ATP5C1,ATP5D,ATP5F1,ATP5H,ATP5I,ATP5J,ATP5O,ATP5S,ATPase,caspase,CCT8,Ck2,F0 ATP synthase,F1 ATPase,FOXRED1,GTPase,HSP90AA1,HSPCA,IARS2,MT-ATP6,Pkc(s),PSMA3,PTCH2,RGS1,RGS11,RNA polymerase II,TPI1,UBB,XAF1,ZNF74
AKIRIN2,ARL17A/ARL17B,C19orf66,C6orf62,CCDC117,CHPT1,CNFN,CSRNP1,EIF4H,ELAVL1,EML4,ETFA,ETFDH,FAM83F,GSTO2,HIGD1A,HNF4A,IER3IP1,KBTBD4,KIAA0247,KLHL6,LGALSL,LIPT1,MRPL57,NSA2,PAQR7,RAB2B,RPL36AL,SLC35A5,STT3A,SZRD1,TMEM68,UBC,UGP2,ZNF106
Akt,Ap1,CD3,COX5B,CYB5A,cytochrome-c oxidase,ERK,estrogen receptor,ETS1,Focal adhesion kinase,Growth hormone,Histone h3,Hsp70,IgG,IL1,IL12 (complex),Immunoglobulin,Integrin,ITGAV,Jnk,LAMB1,LDL,Mapk,MME,NFkB (complex),NRP1,P38 MAPK,PI3K (complex),Pka,PPIAL4G (includes others),SCD,SGK1,Tgf beta,TSPAN3,Vegf
ADCY9,AP5Z1,APP,CCNB1IP1,CHRM4,CRHR2,DENND6B,DNAJC5,DRD2,FAM213A,FOXK1,FOXK2,FYCO1,FZD5,GNB3,GNRHR,GPR3,HCFC2,HTR2A,IRF2BP1,IRF2BP2,IRF2BPL,K Channel,Na+,K+ -ATPase,Na-k-atpase,PAWR,PDXP,PPAP2B,PPP1R1B,PPP3CA,PRDX5,PTGDR,RGS1,SSTR2,voltage-gated calcium channel
NUFIP1,SNORD13

Table S4. Genes correlated with time elapsed in Study I.

Symbol	Probe ID	Rho	P value	D.F.	Adjusted P value
HTRA3	ILMN_1812669	0.758	0.00003	23	0.443
FGFRL1	ILMN_1795865	0.744	0.00005	23	0.443
AGPAT2	ILMN_1732176	0.725	0.00009	23	0.443
TP53I3	ILMN_2358919	0.723	0.00010	23	0.443
PSMB10	ILMN_1683026	0.720	0.00011	23	0.443
RRAGD	ILMN_1699772	0.715	0.00013	23	0.443
G0S2	ILMN_1691846	0.712	0.00014	23	0.443
MYL6	ILMN_2326071	0.707	0.00016	23	0.443
CEBPA	ILMN_1715715	0.707	0.00016	23	0.443
PC	ILMN_1671489	0.695	0.00023	23	0.509
FLJ20254	ILMN_1716907	0.692	0.00026	23	0.509
LGALS1	ILMN_1723978	0.687	0.00030	23	0.509
ADAMTS7	ILMN_2211790	0.683	0.00033	23	0.509
ECHS1	ILMN_1718132	0.683	0.00033	23	0.509
COL5A3	ILMN_1796288	0.681	0.00035	23	0.509
GPX4	ILMN_2378952	0.680	0.00036	23	0.509
DULLARD	ILMN_2133638	0.678	0.00038	23	0.509
PEX19	ILMN_1658759	0.675	0.00041	23	0.509
LETM1	ILMN_1710668	0.673	0.00043	23	0.509
LOC441956	ILMN_1719826	0.673	0.00043	23	0.509
GLYCTK	ILMN_1791222	0.673	0.00044	23	0.509
GLUL	ILMN_1653496	0.669	0.00049	23	0.538
BCR	ILMN_1670398	0.667	0.00051	23	0.538
PKD1L2	ILMN_2372316	0.665	0.00054	23	0.544
NMB	ILMN_2347592	0.662	0.00058	23	0.565
SMPD1	ILMN_1757370	0.656	0.00068	23	0.587
BAI2	ILMN_1773109	0.653	0.00074	23	0.587
SETDB1	ILMN_1718207	0.652	0.00075	23	0.587
MAPK10	ILMN_2340131	0.645	0.00089	23	0.587
ACP6	ILMN_2234343	0.644	0.00091	23	0.587
OSTM1	ILMN_1720303	0.644	0.00092	23	0.587
INPP4A	ILMN_1652647	0.643	0.00094	23	0.587
NOL3	ILMN_2059797	0.643	0.00094	23	0.587
FBXO16	ILMN_1715823	0.640	0.00101	23	0.587
PGM1	ILMN_1800659	0.639	0.00103	23	0.587
CABLES1	ILMN_1653001	0.638	0.00104	23	0.587
FAM90A1	ILMN_1696684	0.636	0.00109	23	0.587
KIAA0182	ILMN_1807767	0.636	0.00109	23	0.587
NCSTN	ILMN_1735180	0.636	0.00111	23	0.587
LOC642946	ILMN_1782178	0.634	0.00116	23	0.587
FAM89A	ILMN_1712859	0.633	0.00117	23	0.587
CNIH3	ILMN_1749071	0.632	0.00120	23	0.587
ICA1	ILMN_1814787	0.632	0.00120	23	0.587
PC	ILMN_2340347	0.632	0.00120	23	0.587
NUTF2	ILMN_1655046	0.630	0.00128	23	0.587

ADAMTS14	ILMN_2358134	0.626	0.00138	23	0.609
PHKG1	ILMN_2113102	0.625	0.00142	23	0.609
AGPAT2	ILMN_1681081	0.624	0.00145	23	0.609
HP1BP3	ILMN_1701169	0.623	0.00150	23	0.619
EPGN	ILMN_1815313	0.618	0.00168	23	0.651
BTD	ILMN_1699728	0.617	0.00172	23	0.651
MYO1C	ILMN_2329165	0.617	0.00172	23	0.651
ZNF423	ILMN_1763602	0.617	0.00172	23	0.651
BMP1	ILMN_1800412	0.612	0.00192	23	0.708
AADAC	ILMN_1760414	0.610	0.00200	23	0.715
DHX9	ILMN_1676285	0.610	0.00200	23	0.715
MSRA	ILMN_2228180	0.609	0.00202	23	0.715
MED10	ILMN_1707631	0.608	0.00207	23	0.721
TNFRSF21	ILMN_1699695	0.604	0.00228	23	0.761
MPV17	ILMN_1691090	0.601	0.00243	23	0.786
POMC	ILMN_2403664	0.600	0.00245	23	0.786
MAPK10	ILMN_1748281	0.600	0.00248	23	0.786
BCL2L13	ILMN_2181445	0.599	0.00253	23	0.792
LGI2	ILMN_1767900	0.595	0.00273	23	0.831
NDUFV2	ILMN_2086417	0.595	0.00273	23	0.831
SLC22A18AS	ILMN_1691048	0.594	0.00281	23	0.837
AQP7	ILMN_1738494	0.591	0.00296	23	0.838
CIDEA	ILMN_2390318	0.591	0.00299	23	0.838
PGA5	ILMN_1717572	0.591	0.00299	23	0.838
MKNK1	ILMN_1750429	0.588	0.00318	23	0.862
CST6	ILMN_1698666	0.587	0.00325	23	0.869
MMP9	ILMN_1796316	0.586	0.00328	23	0.869
SCD	ILMN_1689329	0.583	0.00348	23	0.913
ACADVL	ILMN_2263466	0.581	0.00362	23	0.913
ENO1	ILMN_1710756	0.581	0.00362	23	0.913
FBXL8	ILMN_1682037	0.579	0.00377	23	0.913
APOF	ILMN_1809311	0.577	0.00392	23	0.913
GSS	ILMN_1683462	0.575	0.00412	23	0.913
FCGR2A	ILMN_1706523	0.574	0.00416	23	0.913
KDELR3	ILMN_1798952	0.574	0.00416	23	0.913
RER1	ILMN_1812067	0.574	0.00416	23	0.913
LOC441150	ILMN_1743755	0.573	0.00424	23	0.913
F2RL1	ILMN_1673113	0.573	0.00428	23	0.913
FBLN2	ILMN_1721769	0.573	0.00428	23	0.913
LOC647520	ILMN_1767546	0.573	0.00428	23	0.913
VMO1	ILMN_1735910	0.572	0.00437	23	0.913
WDR79	ILMN_1693669	0.571	0.00441	23	0.913
INF2	ILMN_1727248	0.571	0.00445	23	0.913
LACTB	ILMN_1693830	0.570	0.00449	23	0.913
COPA	ILMN_1811615	0.569	0.00462	23	0.913
LOC653604	ILMN_1793461	0.569	0.00462	23	0.913
UBTD1	ILMN_1794914	0.569	0.00462	23	0.913

GPR64	ILMN_2349071	0.567	0.00476	23	0.913
C1ORF86	ILMN_2097790	0.567	0.00480	23	0.913
IGF1	ILMN_1709613	0.567	0.00480	23	0.913
SPI1	ILMN_2392043	0.566	0.00485	23	0.913
COL1A1	ILMN_1701308	0.566	0.00490	23	0.913
PCOLCE2	ILMN_1746888	0.566	0.00490	23	0.913
SDHB	ILMN_1667257	0.566	0.00490	23	0.913
DBNL	ILMN_2376289	0.565	0.00499	23	0.913
EPM2AIP1	ILMN_1682658	-0.566	0.00490	23	0.913
TMEM178	ILMN_1678403	-0.569	0.00462	23	0.913
TUBB4	ILMN_1682459	-0.571	0.00445	23	0.913
RANBP6	ILMN_1780842	-0.571	0.00441	23	0.913
C3ORF63	ILMN_1661409	-0.572	0.00437	23	0.913
LOC153364	ILMN_1769449	-0.578	0.00385	23	0.913
NOL5A	ILMN_1705407	-0.579	0.00377	23	0.913
LOC391347	ILMN_1654185	-0.588	0.00318	23	0.862
HSPA2	ILMN_1766499	-0.590	0.00302	23	0.838
LRBA	ILMN_1652160	-0.591	0.00299	23	0.838
LOC285053	ILMN_1660832	-0.592	0.00290	23	0.838
LOC374443	ILMN_1708905	-0.594	0.00281	23	0.837
CXCL2	ILMN_1682636	-0.602	0.00235	23	0.776
GOLSYN	ILMN_1738989	-0.605	0.00223	23	0.756
SGK3	ILMN_1747020	-0.606	0.00218	23	0.750
HS.562504	ILMN_1874323	-0.616	0.00173	23	0.651
TWSG1	ILMN_1726967	-0.621	0.00155	23	0.630
LOC647009	ILMN_1739045	-0.624	0.00145	23	0.609
LOC136143	ILMN_1668228	-0.627	0.00137	23	0.609
LOC439949	ILMN_1893633	-0.630	0.00128	23	0.587
HS.545232	ILMN_1875380	-0.631	0.00123	23	0.587
LOC643171	ILMN_1748666	-0.632	0.00122	23	0.587
LOC651453	ILMN_1709948	-0.632	0.00120	23	0.587
FLJ11151	ILMN_1662865	-0.633	0.00117	23	0.587
RND3	ILMN_1759513	-0.640	0.00101	23	0.587
HS.569566	ILMN_1838942	-0.651	0.00078	23	0.587
SMARCA1	ILMN_2376258	-0.654	0.00071	23	0.587

Table S5. Top pathways identified from 116 genes correlated with time elapsed.

Ingenuity Canonical Pathways	p-value
Adipogenesis pathway	0.0009
Mitochondrial Dysfunction	0.0022
Atherosclerosis Signaling	0.0058
Glutamine Biosynthesis I	0.0060
LXR/RXR Activation	0.0060
Hepatic Fibrosis / Hepatic Stellate Cell Activation	0.0066
BMP signaling pathway	0.0100
Intrinsic Prothrombin Activation Pathway	0.0123
Axonal Guidance Signaling	0.0145
Inhibition of Angiogenesis by TSP1	0.0158
Dendritic Cell Maturation	0.0162
Glutathione Biosynthesis	0.0178
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	0.0229
ILK Signaling	0.0234
Biotin-carboxyl Carrier Protein Assembly	0.0240
UVC-Induced MAPK Signaling	0.0263
Gα12/13 Signaling	0.0324
GDP-glucose Biosynthesis	0.0355
RhoA Signaling	0.0355
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	0.0398
Unfolded protein response	0.0407
Glucose and Glucose-1-phosphate Degradation	0.0417
Cardiac Hypertrophy Signaling	0.0427
IL-12 Signaling and Production in Macrophages	0.0447
Glioma Invasiveness Signaling	0.0457
Airway Pathology in Chronic Obstructive Pulmonary Disease	0.0468
Sphingomyelin Metabolism	0.0468
Myc Mediated Apoptosis Signaling	0.0479

Molecules
AGPAT2,ZNF423,CEBPA,SETDB1,FGFRL1
SDHB,NCSTN,GPX4,MAPK10,NDUFV2
MMP9,COL5A3,APOF,COL1A1
GLUL
MMP9,SCD,APOF,ECHS1
MYL6,MMP9,COL5A3,COL1A1,IGF1
MAPK10,BMP1,ZNF423
COL5A3,COL1A1
TUBB4A,MYL6,MMP9,BMP1,MKNK1,ADAMTS7,IGF1
MMP9,MAPK10
MAPK10,COL5A3,COL1A1,FCGR2A
GSS
MAPK10,APOF,SPI1,RND3
MYL6,MMP9,MAPK10,RND3
BTD
MAPK10,SMPD1
MYL6,F2RL1,MAPK10
PGM1
MYL6,RND3,IGF1
MAPK10,BMP1,COL1A1,IGF1
HSPA2,CEBPA
PGM1
MYL6,MAPK10,RND3,IGF1
MAPK10,APOF,SPI1
MMP9,RND3
MMP9
SMPD1
MAPK10,IGF1

Table S6. Top networks identified from 116 genes correlated with time elapsed.

ID	Score	Focus Molecules	Top Diseases and Functions
1	34	18	Hematological System Development and Function, Inflammatory Response, Tissue Development
2	29	16	Developmental Disorder, Hereditary Disorder, Metabolic Disease
3	27	15	Connective Tissue Development and Function, Tissue Morphology, Lipid Metabolism
4	25	14	Dermatological Diseases and Conditions, Developmental Disorder, Hereditary Disorder
5	20	12	Organismal Injury and Abnormalities, Connective Tissue Disorders, Developmental Disorder
6	18	11	Cellular Compromise, Cellular Assembly and Organization, Drug Metabolism
7	18	11	Cancer, Cell-To-Cell Signaling and Interaction, Hematological System Development and Function
8	12	8	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry
9	2	1	Developmental Disorder, Endocrine System Disorders, Gastrointestinal Disease
10	2	1	Cancer, Endocrine System Disorders, Gastrointestinal Disease

Molecules in Network
ACADVL,Akt,AMPK,BCL2L13,BCR,CXCL2,F Actin,FCGR2A,glutathione peroxidase,Glycogen synthase,GOT,GPX4,Ifn,Ige,IgG,IgG1,Igg3,Igm,Immunoglobulin,INPP4A,Interferon alpha,LETM1,LGALS1,MED10,mediator,N-cor,NADPH oxidase,NMB,OSTM1,RND3,SCD,SPI1,TNFRSF21,TUBB4A,TWSG1
ACP6,CLDN12,COPA,ECHS1,ERGIC2,FUNDC2,GPR64,GSE1,KIAA2026,MFSD5,MPV17,NCSTN,NDUFV2,NUP54,NUTF2,PEX3,PEX5,PEX10,PEX12,PEX13,PEX19,PEX26,PEX11B,PGM1,PHKG1,PXMP2,PXMP4,RER1,RRAGD,SACM1L,SLC25A17,TP53I3,UBC,UBTD1,VASN
Actin,AGPAT2,AQP7,CD3,CEBPA,CIDEA,DHX9,ENO1,GLUL,Gsk3,Histone h3,Histone h4,Hsp70,Hsp90,HSPA2,Insulin,Integrin,Jnk,KDELR3,Mapk,MSRA,MYL6,MYO1C,NOP56,P38 MAPK,PC,PI3K (complex),Pka,Pro-inflammatory Cytokine,Proinsulin,Rac,Ras homolog,RNA polymerase II,SETDB1,Vegf
BTD,BTG1,CPPED1,CRNKL1,CTDNEP1,CUL7,EEF2K,EPM2A,EPM2AIP1,FAM208A,FBXL8,FBXL15,FBXO16,GAR1,GLYCTK,HIST2H3D,IFITM3,KDM2A,LPIN2,LRBA,LSM1,MAD2L1BP,MBLAC2,NAF1,NOP10,ORC4,RANBP6,SKP1,SOD2,STK33,TMEM214,TUSC2,UBC,WRAP53,YRDC
20s proteasome,ADAMTS14,Alp,Ap1,APOF,BMP1,C/ebp,COL1A1,COL5A3,collagen,Collagen Alpha1,Collagen type I,Collagen type II,Collagen type III,Collagen type V,Collagen(s),ERK1/2,FBLN2,Fgf,G0S2,gelatinase,Growth hormone,GSS,HDL-cholesterol,IL1,IL-1R,Laminin,LDL,MMP9,PCOLCE2,PDGF BB,PSMB10,SMPD1,STAT5a/b,Tgf beta
Alpha catenin,Beta Arrestin,C8orf44-SGK3/SGK3,caspase,Cg,Creb,cytochrome C,DBNL,E2f,EPGN,ERK,F2RL1,FSH,G protein alphai,Gpcr,HDL,IGF1,INF2,Lh,MAP2K1/2,MAPK10,Mek,MKNK1,NFkB (complex),NOL3,p85 (pik3r),PLC,POMC,Ras,Rock,Sapk,SMARCA1,TCR,Tnf (family),trypsin
AADAC,AQP7,BAMBI,BRD8,CABLES1,CST6,DIAPH3,Enolase,EVPL,EXOC1,EXOC5,FAM90A1,FGF5,FGFR1,FGFRL1,GBP1,GPC1,growth factor receptor,HP1BP3,HSD11B2,HTRA3,ICA1,KLK1,LGMN,LMNA,MYC,NDUFS2,OSM,PGA5 (includes others),RARG,SLC16A3,SRC,TAT,TGM2,ZNF423
ADAMTS7,ADCY7,ASGR1,ATXN1,BAI2,C1orf86,C5AR2,CPB2,CRADD,FANCF,FBXL7,FXN,G0S2,GLS,HCAR3,HNF4A,LACTB,MC3R,NFKBIL1,NPY2R,NPY5R,PEMT,PIK3R5,PPP1R3C,PTGFR,RB1,SDH,SDHB,SLC10A1,SLC22A18AS,SLC52A1,SYBU,TNF,Ubiquitin,ZFP64
PKD1L2,SBDS
ADAM11,ADAM23,LGI2

Table S7. Intrinsic subtype concordance between pairs.

Study I

Sample A \ Sample B	Basal-like	HER2-Enriched	Luminal A	Luminal B	Normal-like
Basal-like	1	0	0	0	0
HER2-Enriched	0	0	0	0	0
Luminal A	0	0	13	**1**	**1**
Luminal B	0	0	**1**	5	0
Normal-like	0	0	0	0	1
Total	1	0	14	6	2

Study II

Diagnosis \ Surgery	Basal-like	HER2-Enriched	Luminal A	Luminal B	Normal-like
Basal-like	0	0	0	0	0
HER2-Enriched	0	0	0	**1**	0
Luminal A	0	0	32	**6**	1
Luminal B	0	0	**4**	10	0
Normal-like	0	0	**2**	0	0
Total	0	0	38	17	1

Total
1
0
15
6
1
23

Total
0
1
39
14
2
56

Table S8. Top pathways identified in Study II.

Canonical Pathways	B-H p-value
IL-17A Signaling in Fibroblasts	0.005
Glucocorticoid Receptor Signaling	0.006
Thrombopoietin Signaling	0.006
CXCR4 Signaling	0.006
ERK5 Signaling	0.006
ILK Signaling	0.006
Prolactin Signaling	0.006
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	0.006
PDGF Signaling	0.006
IGF-1 Signaling	0.010
Colorectal Cancer Metastasis Signaling	0.010
Cholecystokinin/Gastrin-mediated Signaling	0.010
IL-17A Signaling in Gastric Cells	0.011
TNFR2 Signaling	0.012
HMGB1 Signaling	0.012
P2Y Purigenic Receptor Signaling Pathway	0.012
PI3K Signaling in B Lymphocytes	0.013
GNRH Signaling	0.013
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	0.013
Aryl Hydrocarbon Receptor Signaling	0.014
April Mediated Signaling	0.015
MIF Regulation of Innate Immunity	0.016
B Cell Activating Factor Signaling	0.016
UVC-Induced MAPK Signaling	0.017
iNOS Signaling	0.017
TNFR1 Signaling	0.019
Endothelin-1 Signaling	0.019
RAR Activation	0.019
Molecular Mechanisms of Cancer	0.019
CD27 Signaling in Lymphocytes	0.019
NRF2-mediated Oxidative Stress Response	0.019
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	0.019
UVB-Induced MAPK Signaling	0.019
IL-2 Signaling	0.019
IL-8 Signaling	0.019
ERK/MAPK Signaling	0.019
ErbB2-ErbB3 Signaling	0.019
EGF Signaling	0.019
ATM Signaling	0.020
CCR5 Signaling in Macrophages	0.021
Estrogen-Dependent Breast Cancer Signaling	0.021
Pyridoxal 5'-phosphate Salvage Pathway	0.021
CD40 Signaling	0.021
Erythropoietin Signaling	0.021
Neurotrophin/TRK Signaling	0.021

IL-10 Signaling	0.021
GDNF Family Ligand-Receptor Interactions	0.021
Chemokine Signaling	0.021
Renal Cell Carcinoma Signaling	0.022
IL-3 Signaling	0.022
Toll-like Receptor Signaling	0.022
JAK/Stat Signaling	0.022
LPS-stimulated MAPK Signaling	0.022
Signaling by Rho Family GTPases	0.024
Ceramide Signaling	0.026
ErbB Signaling	0.028
RANK Signaling in Osteoclasts	0.028
UVA-Induced MAPK Signaling	0.028
TGF-β Signaling	0.028
PPAR Signaling	0.030
IL-1 Signaling	0.030
T Cell Receptor Signaling	0.031
p53 Signaling	0.032
CDK5 Signaling	0.032
HGF Signaling	0.035
Corticotropin Releasing Hormone Signaling	0.035
Role of Tissue Factor in Cancer	0.035
CD28 Signaling in T Helper Cells	0.035
Renin-Angiotensin Signaling	0.035
PKCθ Signaling in T Lymphocytes	0.035
p38 MAPK Signaling	0.039
14-3-3-mediated Signaling	0.039
IL-6 Signaling	0.039
Cdc42 Signaling	0.044
IL-12 Signaling and Production in Macrophages	0.048
Relaxin Signaling	0.048

Molecules
FOS,JUN,CEBPD
FOS,JUN,DUSP1,SGK1,TSC22D3
MYC,FOS,JUN
FOS,JUN,RHOB,EGR1
MYC,FOS,SGK1
MYC,FOS,JUN,RHOB
MYC,FOS,JUN
FOS,JUN,TOB1
MYC,FOS,JUN
FOS,JUN,CYR61
MYC,FOS,JUN,RHOB
FOS,JUN,RHOB
FOS,JUN
FOS,JUN
FOS,JUN,RHOB
MYC,FOS,JUN
FOS,JUN,ATF3
FOS,JUN,EGR1
MYC,FOS,JUN,CEBPD
MYC,FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
MYC,FOS,JUN
FOS,JUN,DUSP1
MYC,FOS,JUN,RHOB
FOS,JUN
FOS,JUN,JUNB
FOS,JUN,RHOB
FOS,JUN
FOS,JUN
FOS,JUN,RHOB
MYC,FOS,DUSP1
MYC,JUN
FOS,JUN
JUN,GADD45B
FOS,JUN
FOS,JUN
PDXK,SGK1
FOS,JUN
FOS,JUN
FOS,JUN

FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN,RHOB
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
JUN,GADD45B
FOSB,EGR1
FOS,JUN
FOS,JUN
EGR1,CYR61
FOS,JUN
FOS,JUN
FOS,JUN
MYC,DUSP1
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN
FOS,JUN

Table S9. Top networks identified in Stu

ID	Score	Focus Molecules
1	27	12
2	19	9
3	9	5
4	5	3
5	3	2
6	3	1

dy II.

Top Diseases and Functions
Neurological Disease, Cell Death and Survival, Cellular Growth and Proliferation
Endocrine System Disorders, Gastrointestinal Disease, Metabolic Disease
Cell Morphology, Visual System Development and Function, Hereditary Disorder
Gene Expression, RNA Damage and Repair, RNA Post-Transcriptional Modification
Lipid Metabolism, Small Molecule Biochemistry, Drug Metabolism
Cancer, Organismal Injury and Abnormalities, Reproductive System Disease

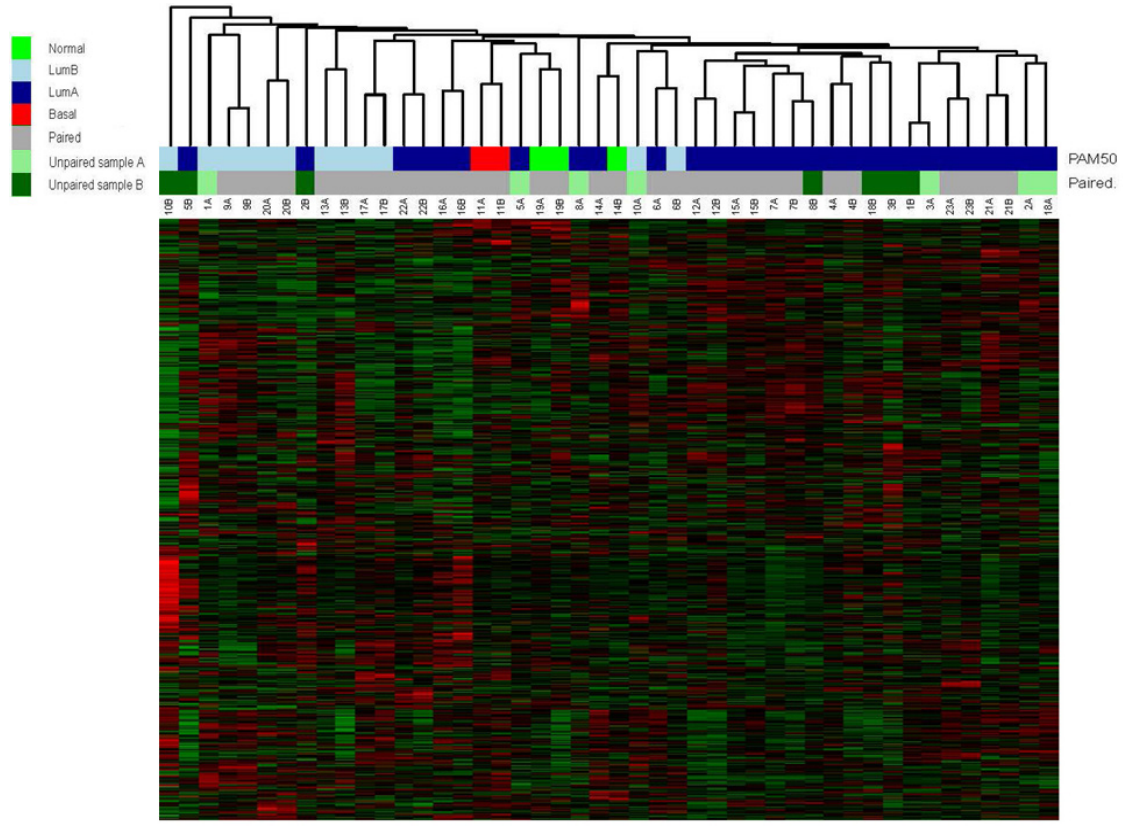
Molecules in Network
Ap1,ATF3,BCR (complex),BHLHE40,C/ebp,Calcineurin protein(s),CaMKII,CCL3L3,CYR61,DUSP1,EGR1,ERK1/2,Fcer1,FOS B,GADD45B,GC-GCR dimer,Gm-csf,HBA1/HBA2,Ige,IL12 (complex),JINK1/2,JUN/JUNB/JUND,JUNB,MAP2K1/2,Nfat (family),PDGF BB,Rar,RASD1,Sapk,SERCA,STAT5a/b,Tgf beta,thymidine kinase,thyroid hormone receptor,TSC22D3
Akt,Alp,BTG2,calpain,Cdc2,CEBPD,Cg,Collagen type I,Creb,Cyclin A,Cyclin E,E2f,ERK,Fgf,FSH,GNRH,Growth hormone,Gsk3,HBB,Hsp27,IL1,Integrin,JUN,LDL,Lh,Mek,Pdgf (complex),PDXK,Pkg,Rb,RGS2,RHOB,Rock,SGK1,TOB1
APOLD1,ARHGEF25,ARID3A,CD163,COPS5,DSG3,FBXL18,FN1,FP R1,IgG,IgG1,Insulin,Jnk,KLK8,KRT13,mir-101,mir-188,miR-532-5p (and other miRNAs w/seed AUGCCUU),NMDA Receptor,P38 MAPK,PDPN,PIP5K1B,Pka,PTPN22,RGS1,RNY5,SERPINB7,SF3A3, SNORD3A,SSB,TGFBI,Tnf (family),TROVE2,VGF,ZNF622
26s Proteasome,ADRB,caspase,CD3,Ck2,Endothelin,estrogen receptor,Focal adhesion kinase,FOS,Gpcr,Hdac,Histone h3,Histone h4,Hsp70,Igm,IKK (complex),Immunoglobulin,Mapk,MYC,NFkB (complex),Nicotinic acetylcholine receptor,Notch,PI3K (complex),Pkc(s),Rac,Ras,Ras homolog,RNA polymerase II,Sos,TCF,TCR,TSH,Ubiquitin,Vegf,ZFP36
ADNP,ALDH1A3,B4GALNT1,CBR3,CTH,GM2A,GOLM1,GSR,HERC 1,HSD17B7,HSD3B2,LECT2,LOC102724428/SIK1,MAN1A2,MAOA ,MC4R,MGST1,MT1A,NDUFA8,NDUFB4,NDUFS7,NEFM,NUCB2,R ABEP2,RGS7,SCO2,Sf1,STAT,SYNPO,TNF,TPP2,UBC,WBSCR22,W NT10B,ZFP36L2
NUFIP1,SNORD13

Table S10. Top 20 genes identified in Study I and their p-value in Study II.

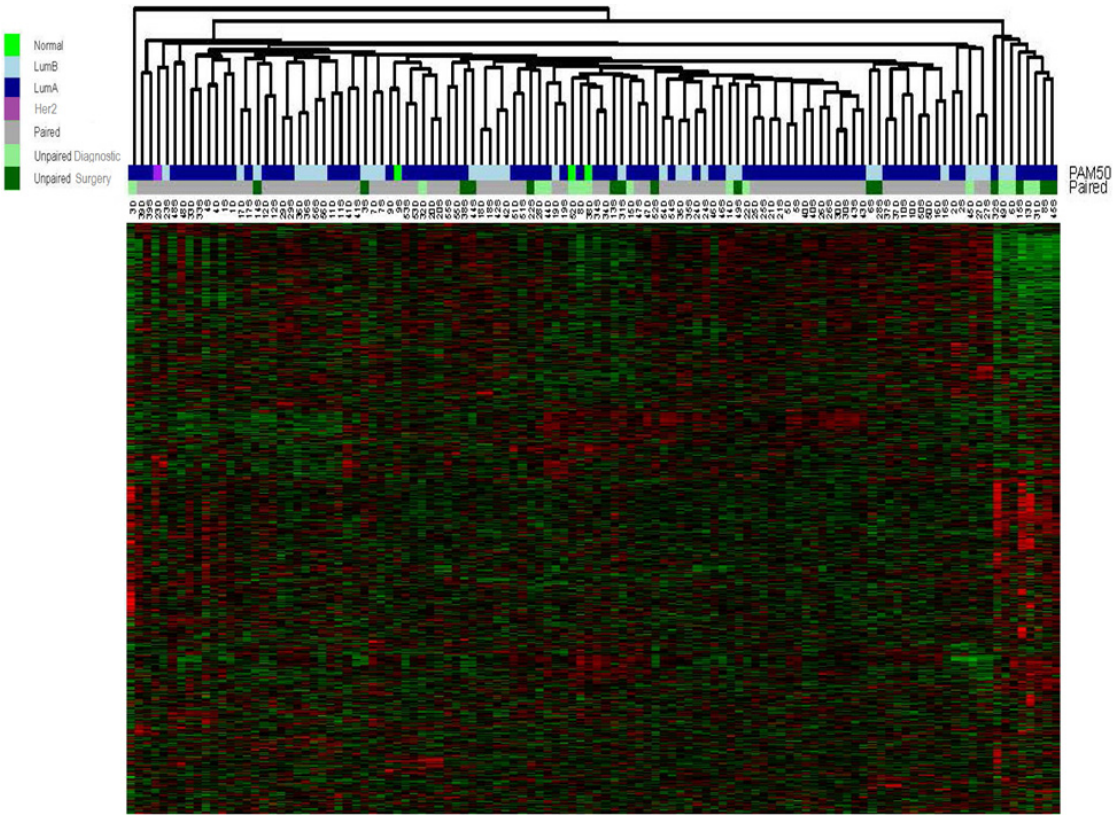
STUDY I						STUD
Accession	**Symbol**	**Parametric p-value**	**FDR**	**FC**		**Parametric p-value**
NM_005252	FOS	< 1e-07	< 1e-07	4.00		0.0144
NM_002922	RGS1	< 1e-07	< 1e-07	3.23		0.0041
NM_004417	DUSP1	< 1e-07	< 1e-07	3.13		0.0003
NM_000517	HBA2	< 1e-05	0.003	-2.90		0.1004
NM_000518	HBB	< 1e-05	0.006	-2.83		0.6704
NM_000517	HBA2	< 1e-05	0.007	-2.64		0.1004
NM_000558	HBA1	< 1e-04	0.008	-2.39		
NM_006732	FOSB	< 1e-06	0.001	2.38		0.0014
NR_001571	RNY5	< 1e-04	0.019	-2.15		
NM_001964	EGR1	< 1e-06	0.001	2.04		0.4480
NM_001554	CYR61	< 1e-05	0.002	2.04		0.0837
NM_003407	ZFP36	< 1e-07	< 1e-07	2.00		0.0005
NR_006882	SNORD3D	< 1e-06	0.001	1.85		
NR_001449	TRK1	< 1e-07	< 1e-07	-1.75		0.1566
NM_002228	JUN	< 1e-07	< 1e-07	1.69		0.0059
NM_005627	SGK	< 1e-06	0.0004	1.64		0.0003
NM_005627	SGK1	< 1e-05	0.002	1.61		0.0003
NR_006881	SNORD3C	< 1e-04	0.010	1.61		
NR_006880	SNORD3A	< 1e-04	0.011	1.61		
NM_005627	SGK1	< 1e-04	0.016	1.56		0.0003

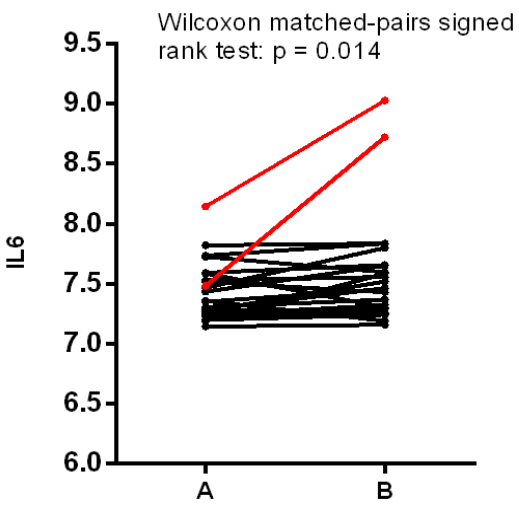
DY II	
FDR	**FC**
0.194	1.64
0.159	1.37
0.133	1.72
0.333	1.23
0.828	1.06
0.333	1.23
0.138	2.08
0.673	1.11
0.312	1.25
0.133	1.54
0.404	-1.05
0.171	1.33
0.133	1.27
0.133	1.27
0.133	1.27

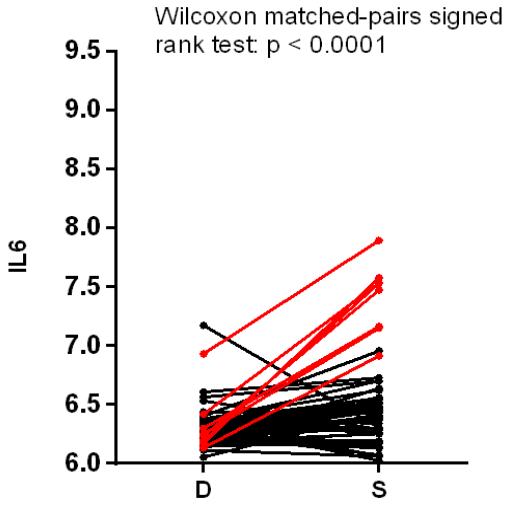
A. Study I



B. Study II




Wilcoxon matched-pairs signed rank test: p = 0.014
IL6
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
A
B


Wilcoxon matched-pairs signed rank test: p < 0.0001
IL6
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
D
S