Cancer Informatics

ORIGINAL RESEARCH

In Silico Discovery of Mitosis Regulation Networks Associated with Early Distant Metastases in Estrogen Receptor Positive Breast Cancers

Yuriy Gusev[1]*, Rebecca B. Riggins[2]*, Krithika Bhuvaneshwar[1†], Robinder Gauba[1†], Laura Sheahan[3], Robert Clarke[2] and Subha Madhavan[1]

[1]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA. [2]Breast Cancer Program, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA. [3]ESAC Inc., Rockville, MD, USA. *,†Equal Contribution.
Corresponding author email: sm696@georgetown.edu

Abstract: The aim of this study was to perform comparative analysis of multiple public datasets of gene expression in order to identify common genes as potential prognostic biomarkers. Additionally, the study sought to identify biological processes and pathways that are most significantly associated with early distant metastases (<5 years) in women with estrogen receptor-positive (ER+) breast tumors. Datasets from three published studies were selected for in silico analysis of gene expression profiles of ER+ breast cancer, using time to distant metastasis as the clinical endpoint. A subset of 44 differently expressed genes (DEGs) was found common to all three studies and characterized by mitotic checkpoint genes and pathways that regulate mitotic spindle and chromosome dynamics. DEG promoter regions were enriched with NFY binding sites. Analysis of miRNA target sites identified significant enrichment of miR-192, miR-193B, and miR-16-1 targets. Aberrant mitotic regulation could drive increased genomic instability leading to a progression towards an early onset metastatic phenotype. The relative importance of mitotic instability may reflect the clinical utility of mitotic poisons in metastatic breast cancer, including poisons such as the taxanes, epothilones, and vinca alkaloids.

Keywords: estrogen receptor alpha-positive, mitotic checkpoint signaling, mitotic regulation network, microRNA targets, early distant metastasis

Cancer Informatics 2013:12 31–51

doi: 10.4137/CIN.S10329

This article is available from http://www.la-press.com.

Introduction

As the second-leading cause of cancer-related death in women, breast cancer is a serious public health problem. The American Cancer Society reported over 228,000 new cases of breast cancer in 2012 and nearly 40,000 breast cancer related deaths in women per year in the United States.[1] A significant majority (~70%) of these women have a tumor that is estrogen receptor alpha-positive (ER+).[2] While we have made significant progress in our understanding of primary breast tumors, we lack sufficient knowledge and understanding of distant metastasis.[3] Gene expression analysis and other molecular profiling tools are revolutionizing our understanding of all cancers.[4] Multiple groups have used these technologies to construct specific gene signatures that aim to identify patients with ER+ breast cancer and thus who are at risk for developing resistance to endocrine or anti-hormonal therapy.[5–8]

However, the accuracy, sensitivity, and specificity of these signatures have been difficult to validate in follow-up studies.[9] Tumor sample and patient population heterogeneity could certainly be confounding factors, but arguably the main complication is the unique challenge of analyzing large-scale, high-dimensional genomic and/or proteomic metadata in a biologically meaningful fashion.[10] Most importantly, beyond the identification of classifiers, the field lacks a clear path along which to move or progress. For these profiles and signatures to be truly useful to the thousands of women diagnosed with ER+ breast cancer every year, we must also use them to understand cancer biology at the functional (mechanistic) and phenotype levels and to inform the design of novel therapeutic agents.

There are several reasons for revisiting published studies of gene expression profiling in cancer and for conducting additional comparative analyses. Global profiling of gene expression provides comprehensive information on genome-wide transcriptome alterations in tumor samples and any individual study rarely utilizes all available measurements. This rich resource of publicly available data is simply not being fully utilized for downstream analysis. It is also increasingly evident that the specific gene expression signatures reported in many published studies are not reproducible.[10] However, the underlying biological processes, pathways, and functional categories by which these genes are annotated appear to be much more common; therefore, downstream systems biology analysis is becoming increasingly more important. Fortunately, computational tools and functional annotation resources are constantly being improved and updated, providing new opportunities for deeper analysis and biological interpretation of expression profiling results.

Over the past three years we have developed a Georgetown Database of Cancer (G-DOC) where many public and clinically relevant molecular profiling studies are being stored. The database has a major focus on breast cancer.[11] G-DOC is a publicly available Web platform that enables translational cancer research by integrating patient characteristics and clinical outcome data with a variety of high-throughput research data, all in a unified environment. G-DOC includes a broad collection of bioinformatics and systems biology tools for the analysis and visualization of four major "omics" types: DNA, mRNA, microRNA, and metabolites. By establishing a standard uniform data processing pipeline and robust quality control of the data, we have accumulated more than 20 breast cancer studies that are processed in a similar manner, allowing for cross study comparisons and further analysis. Currently, G-DOC contains data from more than 3600 breast cancer cases and 1700 gastrointestinal cancer cases. The three studies presented here are also available as part of the G-DOC collection.

Given the prevalence (~70% of cases diagnosed annually),[2] unique biology, and long-term recurrence profile of ER+ breast cancer, we focused our efforts on the development of prognostic biomarkers of early distant metastases (<5 years) in women with this type of breast cancer. After breast-conserving surgery and appropriate radio and systemic therapy (endocrine and/or conventional chemotherapy), the risk of distant metastasis in ER+ breast cancer patients peaks between 2 and 3 years post-diagnosis.[12] Risk gradually decreases after this point but reaches a plateau at 5 years; no further reduction in risk is observed, even at 10 years post-diagnosis. This is in stark contrast to women with ER– breast tumors; these women show a much higher risk of distant metastasis at 2 years post-diagnosis, but from 5 to 10 years post-diagnosis (and presumably, beyond) are actually less likely to present with distant metastasis when compared to their counterparts that have ER+ breast cancer. Mortality rates for ER+ and ER– breast cancer are broadly consistent

with this trend; risk of death for women with ER– breast cancer becomes lower than that for women with ER+ breast cancer at 6 years post-diagnosis.[12] Given these data, we proposed two hypotheses. Firstly, we hypothesize that primary tumors from women with ER+ breast cancer who go on to develop early distant metastasis are characterized by molecular functions and biological processes distinct from those in women with ER+ breast cancer who do not. Secondly, we hypothesize that key components of these signaling networks are amenable to established or emerging targeted therapeutic agents. Our discovery of a common network of overexpressed and highly interconnected genes controlling the entire sequence of mitotic events in the three studies we examined of patients with recurrence within 5 years of initial diagnosis provides insight into potential large scale deregulation of the mitotic machinery that is associated with, and might even facilitate, a progression toward early metastasis.

Methods

Study selection

Datasets from three previously published studies (described below) were selected for in silico identification of a gene expression signature for ER+ breast cancer, early distant metastasis. Patient groups were defined as either (a) ER+ breast cancer patients with no documented distant metastasis at >5 years (Group 1, no met) or (b) ER+ breast cancer patients with documented distant metastasis at ≤5 years (Group 2, met). All data were derived from surgical specimens of ER+ primary breast tumors arrayed on Affymetrix U133A GeneChips.

Patient groups consisted of the Loi dataset (two studies), the Sotiriou dataset, and the Desmedt dataset. Additionally, the Schmidt dataset acted as a Validation Set.

The Loi datasets consisted of two studies.[8] Set A consisted of 255 patients with early-stage (I,II) ER+ and/or progesterone receptor positive (PR+) breast cancer that received either Tamoxifen or no systemic adjuvant treatment. Set B consisted of 355 patients with early-stage (I,II) ER+ breast cancer that received either Tamoxifen or no systemic adjuvant treatment. Raw data were obtained from GEO (accession number GSE6532). Patients in both studies/datasets had a mix of lymph node-negative and -positive disease.

The Sotiriou dataset[13] dataset consisted of samples from 189 patients with primary operable invasive breast cancer. The training set KJX64 contained data from 64 women with ER+ primary breast cancer who received either Tamoxifen or no systemic adjuvant treatment; the validation set KJ125 contained data from 125 ER+ and ER– breast tumor samples. Only ER+ patients were used in our analysis. No patient in the KJ125 validation set received any adjuvant systemic therapy, and patients in both sets had a mix of lymph node-negative and -positive disease. Raw data were obtained from GEO (accession number GSE2990).

The Desmedt dataset[14] consisted of 198 samples from women with lymph node-negative ER+ and ER– breast cancer who received no systemic therapy; only ER+ patients were used in our analysis. Raw data were obtained from GEO (accession number GSE7390).

Upon completion of our comparative analysis of 3 breast cancer studies we identified an additional published breast cancer dataset that was used as a "validation" dataset (Schmidt dataset).[15] The main purpose of this additional analysis was to check if other studies based on distant metastasis outcome contain similar patterns of gene expression and if similar functional groups and pathways could be identified as most significantly affected in this additional dataset. This additional dataset consisted of 200 samples from women with lymph node-negative ER+ breast cancer who received no systemic therapy. Raw data were obtained from GEO (accession number GSE11121).

Data processing and gene selection

A data processing pipeline was established for uniform processing, normalization, and in silico analysis of raw gene expression data. Data normalization and quality control scripts were written in R with tools from the open source software package Bioconductor. Data processing for each dataset was conducted using this standard pipeline, including background correction, normalization with RMA (Robust Multichip Average),[16] and Median Polish[17] followed by robust quality control. Group comparison was performed using LIMMA[18] with multiple testing correction based on Benjamini and Hochberg false discovery rate estimates (Bioconductor).[19] From each dataset, we identified differentially expressed genes with a

greater than 1.0 fold change and adjusted *P*-values $P \leq 0.05$.

Systems biology analysis

These gene sets were further analyzed using several systems biology tools in order to obtain biologically relevant functional annotations. The systems biology analysis pipeline included a combination of original methods developed by our team, open source software, and proprietary software. For each study, we conducted functional profiling using Gene Ontology Enrichment, Pathway Analysis Enrichment using multiple pathway databases, Gene Set Enrichment Analysis, and Regulatory Subnetwork Enrichment (Pathway Studio 9), as well as upstream and downstream common regulators analysis (Pathway Studio 9). These methods were also applied to a list of DEGs common to all three studies. This common subset of DEGs was further analyzed using network analysis based on known protein-protein interaction databases (Reactome,[20] STRING[21]), as well as Ingenuity Knowledgebase and Ingenuity Pathways Analysis (IPA) 8.6 (Ingenuity Systems, www.ingenuity.com). Parameter settings for all of the enrichment analyses—including GO ontology enrichment, pathway enrichment, and subnetwork enrichment—were the same with *P*-values generated based on Fisher's exact test and using a *P*-value threshold of 0.05 ($P \leq 0.05$).

Statistical analysis

Statistical analysis of breast cancer clinical data was performed using Prism 5.0c for Mac (GraphPad Software, San Diego, CA). Survival functions for all 3 studies were estimated by Kaplan-Meier analysis and compared using the Mantel-Cox log rank test. For the Loi and Sotiriou datasets, differences in the distribution of lymph node (LN) status (LN+ vs. LN–) and adjuvant therapy (Tamoxifen treatment vs. no systemic therapy) between Groups 1 and 2 were assessed by Fisher's exact test or χ^2 analysis; this analysis was not performed on the Desmedt dataset because all patients were untreated and LN–. Differences in tumor size and age between Groups 1 and 2 were determined by two-tailed, unpaired *t*-test (parametric) or Mann-Whitney rank sum test (non-parametric), as appropriate. For all three studies, $P \leq 0.05$ was considered statistically significant.

Results

Comparative analysis of three breast cancer studies

To identify differentially expressed genes that are characteristic of early distant metastasis in ER+ breast cancer, we selected three publicly available datasets with sufficient length of follow-up to perform a meaningful comparison between those patients with documented distant metastasis within 5 years and those with no distant metastasis within 5 years. The Loi, Sotiriou, and Desmedt studies[8,13,14] fulfill this criterion and show no significant difference in overall distant metastasis (DM)-free survival proportions (Additional file 1). To identify potential sources of bias between patient groups, we examined four clinical parameters that could serve as independent poor prognostic factors for distant metastasis, independent of gene expression signature(s): primary tumor size, age, lymph node status, and adjuvant therapy.[22]

Age, nodal status, and the distribution of Tamoxifen-treated and untreated patients are equivalent (data not shown), but primary tumor size was significantly larger in Group 2 patients (distant metastasis within 5 years) than in Group 1 patients (no distant metastasis after 5 years) for all three studies (Additional file 2).

Enrichment analysis of differentially expressed genes indicates common mitosis-related processes affected in all 3 studies

We next identified statistically significant, differentially expressed genes (DEGs; fold change > 1.0 and $P \leq 0.05$) for each of the three studies (Additional file 3). For each set of DEGs, functional profiling using Gene Ontology Enrichment was performed. Strikingly, many of the top-ranked functional categories from each study are related to specific mitotic events and mechanisms. For example, in the Loi study these categories included kinetochore and spindle assembly processes, checkpoint regulation, spindle organization, positive regulation of exit from mitosis, and other categories related to mitotic events. We then used Gene Set Enrichment Analysis (GSEA) to query a large collection of databases including Reactome, Biocarta, KEGG, and Pathway Commons; we identified additional functional categories relevant to cell cycle checkpoints. In particular, GSEA results for the

Loi study include the role of Ran in mitotic spindle regulation and genes involved in prometaphase (Table 1). When compared, these three studies had the same top-ranked GO categories statistically significantly enriched for cellular processes associated with mitosis (Fig. 1). More detailed analyses were performed on a subset of 44 DEGs (39 up-regulated, 5 down-regulated) common to all three studies (intersection gene set, Table 2). The expression pattern of this intersection gene set (IGS) was consistent across all 3 studies with almost all 44 genes showing overexpression in early metastasis group (Fig. 2). These 44 genes were functionally annotated using Ingenuity Knowledge Base (Additional file 4). For almost all of the genes, we found that their annotations include associations with mitotic checkpoints and/or specific mitotic events. However, since these genes exhibit similar patterns of co-expression in all three studies, it is of considerable interest to analyze these genes as a group to better understand the underlying subnetwork structure of this gene set and their collective involvement in regulation of specific mitotic events.

Pathway and subnetwork enrichment analysis using multiple systems biology tools

We took several different approaches to both better understand the functional commonalities between members of the IGS and to determine which gene changes are correlated with specific molecular events. The first was to conduct a comprehensive functional annotation and profiling of this cluster of differentially expressed genes by applying enrichment analysis similar to the analysis used for each dataset. We have extended this type of analysis to a variety of experimentally derived functional categories of genes/proteins such as GO biological processes and a common collection of canonical signaling and regulatory pathways. Using GO enrichment analysis, we found a large number of mitosis related processes that were significantly enriched with this 44 gene IGS. The top 20 categories were all related to specific mitotic events/processes and are shown in Table 3. Additional enrichment analysis of IGS was conducted using two independent commercial knowledge bases developed by Ariadne Genomics and Ingenuity Systems Inc. The results obtained with both knowledge bases were consistent with each other and similar to previous GO enrichment analysis. Due to the difference in content of these two knowledge bases, some pathways are present in only one of the databases. For example, the Ingenuity analysis showed that a top ranked pathway was Polo-like kinase 1 (PLK1) in regulation of mitosis (Fig. 3), while it is not present in the Pathway Studio pathway collection. Overall, we found multiple top ranked categories related to biological function and regulatory pathways of mitotic checkpoints using both Ariadne and Ingenuity databases.

Subnetwork enrichment analysis (pathway studio)

The IGS showed very similar clusters of overexpressed genes downstream of a small number of key

Table 1. Gene set enrichment analysis results for Loi studies.

Gene set name	Description	# genes in overlap	*P*-value
REACTOME_CELL_CYCLE_MITOTIC	Genes involved in cell cycle, mitotic	56	$4.45\ e^{-6}$
REACTOME_MITOTIC_M_M_G1_PHASES	Genes involved in mitotic M-M/G1 phases	33	$2.57\ e^{-5}$
REACTOME_MITOTIC_PROMETAPHASE	Genes involved in mitotic prometaphase	22	$7.84\ e^{-5}$
REACTOME_G2_M_CHECKPOINTS	Genes involved in G2/M checkpoints	13	$1.95\ e^{-4}$
REACTOME_E2F_MEDIATED_ REGULATION_OF_ DN_DNA_REPLICATION	Genes involved in E2F mediated regulation of DNA replication	11	$2.32\ e^{-4}$
KEGG_CELL_CYCLE	Cell cycle	26	$3.25\ e^{-4}$
REACTOME_E2F_TRANSCRIPTIONAL_ TARGETS_ AS_AT_G1_S	Genes involved in E2F transcriptional targets at G1/S	8	$8.71\ e^{-4}$
REACTOME_G1_S_TRANSITION	Genes involved in G1/S transition	21	$9.9\ e^{-4}$
REACTOME_CELL_CYCLE_CHECKPOINTS	Genes involved in cell cycle checkpoints	22	$1.14\ e^{-3}$
BIOCARTA_RANMS_PATHWAY	Role of Ran in mitotic spindle regulation	5	$1.64\ e^{-3}$

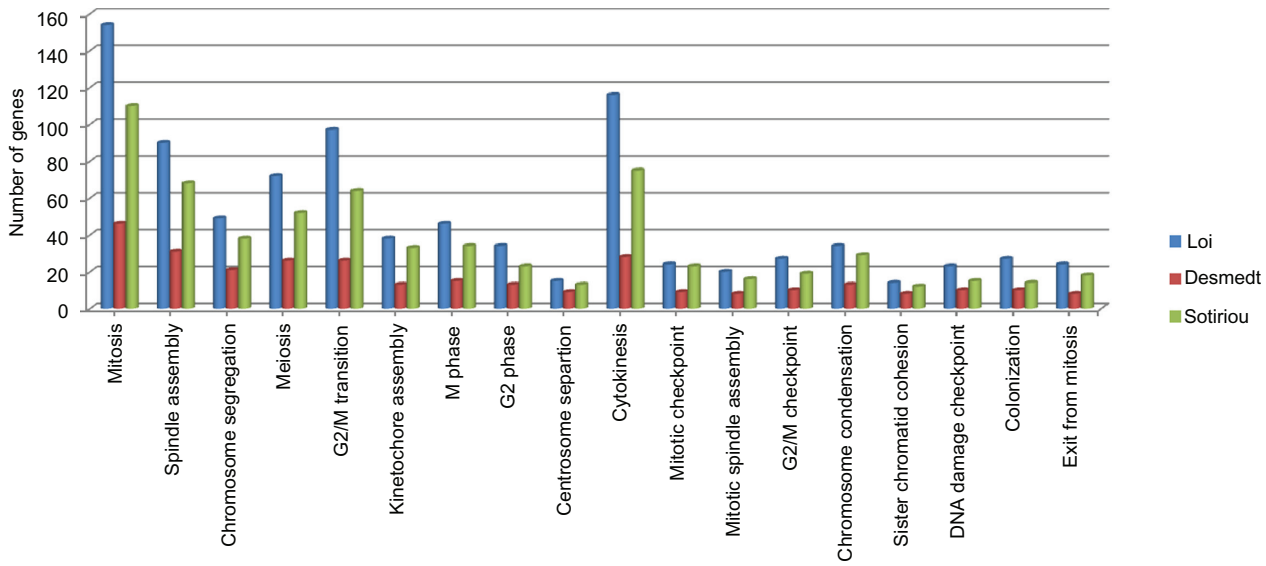

Figure 1. Results of GO enrichment analysis showing top-ranked biological processes enriched with DEGs for the three studies.

regulators of mitosis. For example, in each study PLK1 and CCNB1 (cyclin B1) nucleate a series of highly overlapping subnetworks of uniformly over-expressed genes (Fig. 4). Analysis of subnetwork enrichment of gene sets associated with known cell processes (Pathway Studio) has shown similar results with all top ranked processes related to specific mitosis-related processes (Additional file 5). For instance gene sets from the three top ranked cellular processes (spindle assembly, chromosome segregation, and genome instability) were significantly enriched with genes from the IGS (Fig. 5).

Transcription factor regulation

Additional analysis was performed to determine whether there are functional modules among these 44 genes that are sharing common transcription factor (TF) regulation. TF regulation was determined by applying transcription factor analysis in Ingenuity IPA and subnetwork analysis of common expression regulators (Pathway Studio, Ariadne). Several TFs were identified as having their target genes enriched with genes from the intersection gene set. The top 10 TFs ranked by significance of subnetwork enrichment include E2F4 as the top ranked regulator of expression, followed by FOXM1 (Table 4). Additional promoter enrichment analysis performed by directly querying Molecular Signatures Database (MSigDB)[23] indicated that promoter regions of 14 of the 44 genes contain at least one consensus binding site for the transcription factor NFY ($P = 0.009$); serum response factor (SRF) and E2F sites were also significantly enriched ($P < 0.05$).

microRNA target gene analysis

Analysis was done to determine groups of genes from IGS that are collectively targeted by specific microRNAs. The IGS of 44 genes was analyzed using a similar subnetwork enrichment methodology applied to subnetworks of microRNA target genes. An intersection of target predictions by at least two of three prediction algorithms (TargetScan, Miranda and Pictar)—combined with a large set of experimentally validated targets—was used to identify subnetworks of co-regulated genes in PS Ariadne. A large number of additional experimentally validated targets were derived from Tarbase 6.0,[24] uploaded into Pathway Studio, and included in the enrichment analysis together with computationally predicted targets. Several connected subnetworks of microRNAs and their targets were identified with 29 genes targeted by 12 microRNAs (Fig. 6A), which represents a highly interconnected mitotic regulation network of miRNA and their targets. Importantly, a majority of microRNA-target interactions in this network were identified as experimentally validated according to Tarbase 6.0 with supporting evidence from published studies. The exceptions are for 3 interactions: miR-132 with BRCA1; miR-27B with NEK2; miR-98

Table 2. List of 44 genes common to all three studies with expression change indicated up/down.

Loi_Desmedt_Sotiriou	Expression
AURKA	Up
BRCA1	Up
BUB1B	Up
CCNB1	Up
CCNB2	Up
CDC25C	Up
CDCA3	Up
CDKN3	Up
CENPA	Up
CSE1L	Up
DIXDC1	Down
DSCC1	Up
DUSP4	Down
ESPL1	Up
FANCI	Up
GINS2	Up
GNG12	Down
HMMR	Up
HNRNPAB	Up
KIF15	Up
KIF18A	Up
KIF18B	Up
KIF4A	Up
KIFC1	Up
LRRC59	Up
MCM2	Up
MELK	Up
NEK2	Up
PDLIM4	Down
PLK1	Up
PRC1	Up
PTTG1	Up
RACGAP1	Up
SPAG5	Up
TIMELESS	Up
TMEM106C	Up
TMPO	Up
TOP2A	Up
TPX2	Up
TRIP13	Up
TROAP	Up
UBE2C	Up
UBE2S	Up
ZFP36L2	Down

with DUSP4. These 3 predicted interactions are shown as dotted lines in grey color on Figure 6A. Noticeably, BRCA1 is targeted by most the microRNAs (eight microRNAs; all but one (miR-132) are experimentally validated); miR-192 targets the largest number of genes from IGS (11 genes, all interactions are experimentally validated according to Tarbase 6.0).

Common transcription regulators of miRNAs and their targets

Since microRNA and their target genes might have some common upstream regulators of expression, an additional round of enrichment analysis was performed to determine common transcription factors that could control expression of this subnetwork of microRNAs and their targets. This analysis was based on experimental evidence collected within the Resnet database (Pathway Studio). The top ranked TF was TP53, which has known and validated interactions with several genes and microRNAs from our regulatory module. The top three TFs appear to regulate a significant part of the miRNA-gene mitotic regulation network (Fig. 6B, TFs are highlighted in blue, edges with arrows indicate transcriptional activation, edges with blunt ends indicate transcriptional inhibition).

Protein interaction analysis

We next sought to understand how the protein products of these IGS genes might physically and/or functionally interact, using Reactome and STRING knowledge bases. Consistent with other analyses, Reactome revealed significant enrichment in multiple categories related to mitotic signaling, genome/chromosome stability, and DNA replication and repair (Additional file 6), with an even greater number of mitotic events enriched with genes from a 44 gene intersection set (75 categories out of 81). Using STRING we identified a large association network with highly interconnected nodes, including all but seven proteins encoded by members of the 44-gene IGS. The large fraction of associations between the proteins is based on evidence of co-expression or co-occurrence, while a smaller subnetwork of connections linking ~25% of the genes in our intersection list is based on direct protein-protein interactions, including CCNB1, PLK1, AURKA, and UBE2C (Fig. 7A). Using STRING functionality, we generated a protein co-expression matrix for the entire set of 44 genes (Fig. 7B). Importantly, co-expression evidence collected in the STRING knowledgebase for homo sapiens shows that the majority of proteins are co-expressed with high confidence across multiple experimental settings and various human tissues. Co-expression matrices also show discrete clusters of co-expressed proteins implying several co-regulated, functional modules (Fig. 7B).

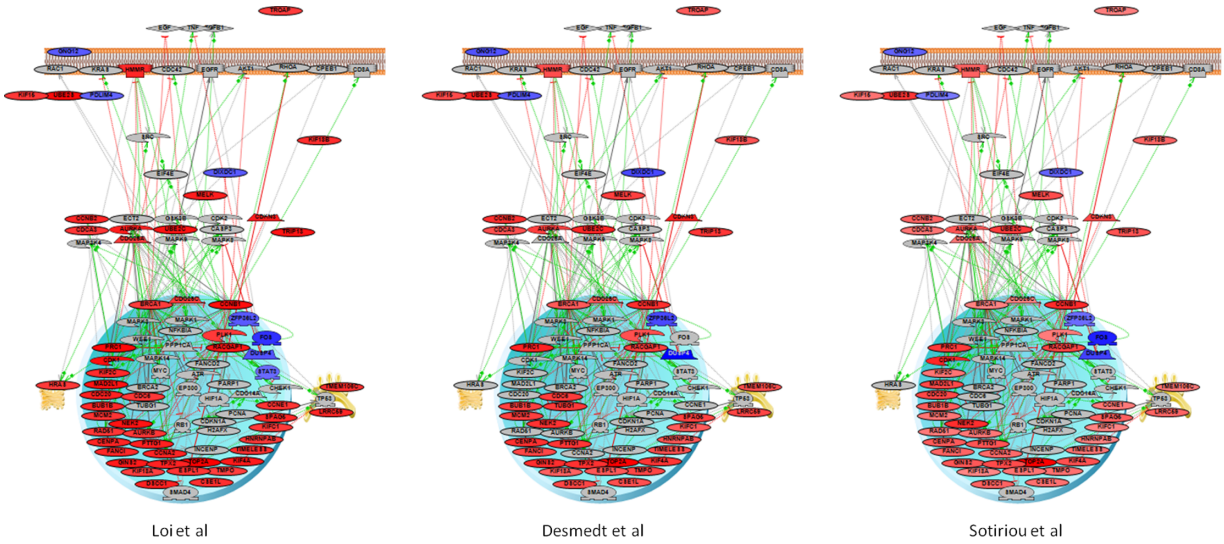

Figure 2. Pathway diagrams of intersection DEGs for each of the three studies.
Notes: Color of the nodes corresponds to a level of expression as fold change. Red—overexpression; Blue—downregulation, Grey nodes represent other interacting proteins with no significant change.

Downstream regulation network analysis

Based on the enrichment analysis of common regulators of downstream cellular processes, it is evident that the 44 differentially expressed genes are involved with regulation of an unusually large number of specific biological events and processes that are intrinsic to mitosis. Cell progression through the consecutive stages of mitosis involves a sophisticated and redundant network of highly interactive genes. Currently, this elaborate network of regulatory interactions is not well understood and poorly annotated at the level of pathways.

Using computational tools provided by Pathway Studio we combined our enrichment findings for individual cellular processes and attempted to construct a crude, granular level representation of mitotic pathways that include approximately 20 specific mitotic events. These events include the initiation of mitosis, multiple intermediate steps associated with mitotic spindle formation, and final steps of chromosome segregation and cytokinesis (Fig. 8). Twenty-nine genes from the intersection gene set were found to be directly associated with at least two mitotic processes. We have conditionally grouped these genes in two categories: major mitotic regulators that are involved with more than five mitotic processes, and those with five or less. Interestingly, by overlaying our findings regarding enrichment of disease phenotypes (presented below), we found that 9 out of 10 metastasis-associated genes are in the major mitotic regulators category (Fig. 8, nodes are highlighted with green color).

To understand cellular processes correlated with disease phenotype, a subnetwork analysis was done for enrichment of the disease phenotype gene sets (Ingenuity, Ariadne). We found that all significantly enriched disease-related gene sets represent phenotypes associated with cancer in general; these are also associated with several specific types of cancer that includes breast cancer and colorectal cancer (Fig. 9). Interestingly, the top three ranked categories were aneuploidy, polyploidy, and tetraploidy, a fact supported by a large number of published evidence that connects 20 out of 44 genes to ploidy related phenotypes. One of the significantly enriched disease categories was related to a phenotype of metastasis with 10 out of the 44 genes overlapping in this category (Fig. 9, represented by green circles). All of these genes belong to highly interconnected hubs within the 44 gene network and all are directly involved with regulation of multiple mitotic events.

We have also examined the intersection DEG list for genes whose protein products are implicated in response to either conventional cytotoxic or novel targeted therapeutic agents. TOP2A (target of epirubicin),[25] PLK1 (target of BI 2536),[26,27] and AURKA (target of MLN8054 and MLN8237)[28,29] are each overexpressed in primary tumors from patients

Table 3. Top 20 biological processes as a result of GO enrichment analysis for 44 DEGs from intersection set.

Name	Total entities	Overlap	Percent overlap	Overlapping entities	*P*-value
Cell cycle	604	23	3	PLK1, PTTG1, BRCA1, CDC25C, BUB1B, CCNB2, AURKA, UBE2S, TPX2, SPAG5, UBE2C, KIFC1, MCM2, RACGAP1, PRC1, NEK2, FANCI, TIMELESS, DSCC1, CDCA3, CDKN3, DIXDC1, KIF18B	5.23E-29
Cell division	336	18	5	PLK1, PTTG1, CCNB1, CDC25C, BUB1B, CCNB2, AURKA, UBE2S, TPX2, SPAG5, UBE2C, KIFC1, RACGAP1, PRC1, NEK2, TIMELESS, CDCA3, KIF18B	4.96E-25
Mitosis	252	15	5	PLK1, PTTG1, CDC25C, BUB1B, CCNB2, AURKA, TPX2, SPAG5, UBE 2C, KIFC1, NEK2, TIMELESS, CDCA3, KIF15, KIF18B	1.72E-21
Mitotic cell cycle	316	14	4	PLK1, PTTG1, CCNB1, CDC25C, BUB1B, CCNB2, AURKA, UBE2C, MCM2, CENPA, NEK2, GINS2, KIF18A, KIF18B	2.98E-18
Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	90	7	7	PLK1, PTTG1, CCNB1, BUB1B, AURKA, UBE2S, UBE2C	2.99E-11
Microtubule-based movement	121	6	4	KIFC1, RACGAP1, KIF4A, KIF18A, KIF15, KIF18B	1.31E-08
Spindle organization	22	4	18	BUB1B, AURKA, SPAG5, UBE2C	1.95E-08
Cell cycle checkpoint	141	6	4	CCNB1, CDC25C, BUB1B, CCNB2, UBE2C, MCM2	3.28E-08
Chromosome segregation	75	5	6	PTTG1, ESPL1, BRCA1, TOP2A, NEK2	5.26E-08
Phosphatidylinositol-mediated signaling	77	5	6	BUB1B, AURKA, SPAG5, UBE2C, TOP2A	6.00E-08
Mitotic prometaphase	90	5	5	PLK1, CCNB1, BUB1B, CENPA, KIF18A	1.32E-07
DNA replication	179	6	3	BRCA1, CDC25C, MCM2, TOP2A, GINS2, DSCC1	1.35E-07
M phase of mitotic cell cycle	96	5	5	PLK1, CDC25C, BUB1B, CENPA, KIF18A	1.82E-07
G2-M transition of mitotic cell cycle	116	5	4	PLK1, CCNB1, CDC25C, CCNB2, NEK2	4.69E-07
Mitotic sister chromatid segregation	14	3	21	ESPL1, KIFC1, NEK2	7.91E-07
Cytokinesis	60	4	6	PLK1, ESPL1, RACGAP1, PRC1	1.25E-06
Cell proliferation	429	7	1	PLK1, CDC25C, BUB1B, TPX2, ZFP36L2, CSE1L, KIF15	1.46E-06
Free ubiquitin chain polymerization	2	2	100	UBE2S, UBE2C	1.73E-06
Regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	77	4	5	PLK1, CCNB1, BUB1B, UBE2C	3.42E-06
Homologous chromosome segregation	4	2	50	PTTG1, ESPL1	1.04E-05

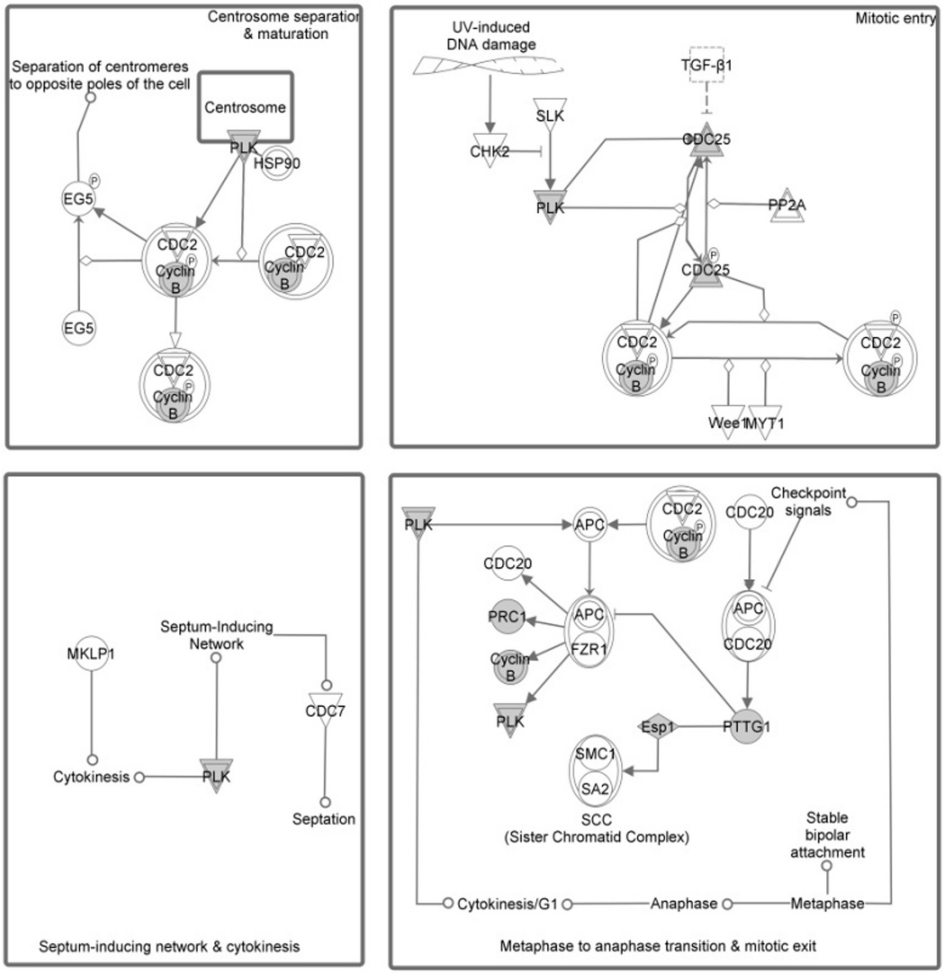

Figure 3. Results of pathway enrichment analysis (Ingenuity Pathway Analysis 8.5) for 44 intersection genes from the three studies.
Note: Top ranked significantly enriched group shows four pathways related to Polo-like Kinase 1 regulation of mitotic events.

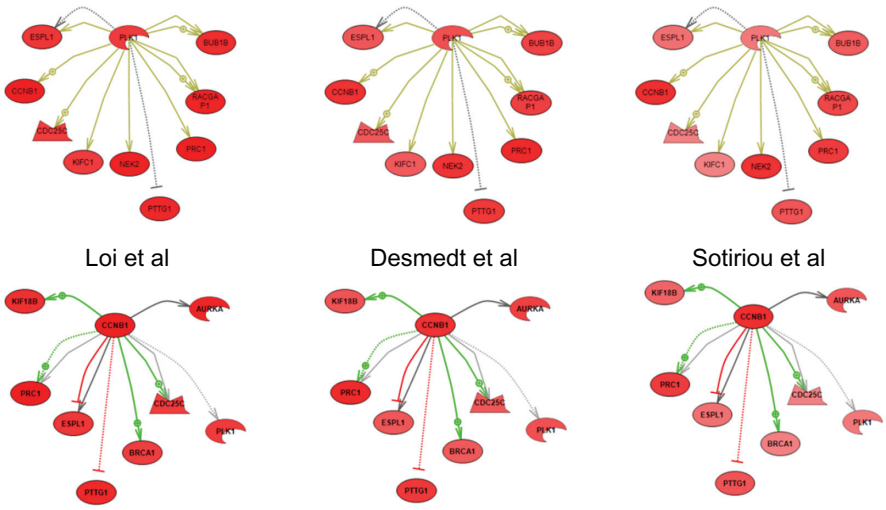

Figure 4. Subnetwork enrichment analysis results for upstream regulators of expression of 44 intersection genes.
Notes: Top two regulators are shown as nodes with downstream interactions indicated by edges. Colors of the nodes correspond to a level of expression as fold change. Red—overexpression. Color of the edges corresponds to a type of regulation of expression: Green—known positive regulation; Red—known negative regulation; Dark Grey—direct regulation with unknown effect; Light Grey—general regulation with unknown effect; Yellow—protein modification; Arrows indicate upregulation; Blunt ends indicate downregulation.

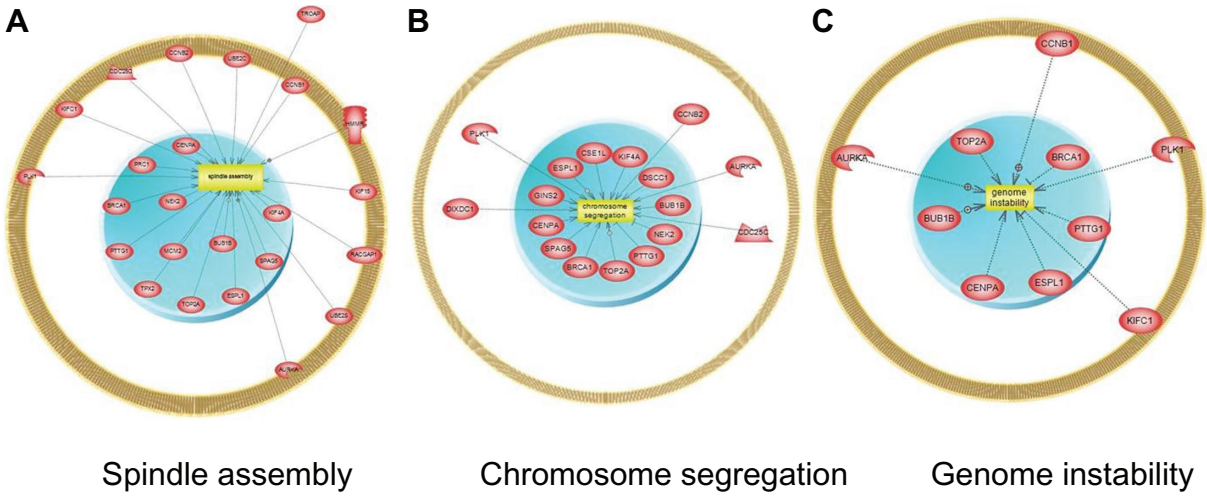

Figure 5. Subnetwork enrichment analysis of downstream cellular processes regulated by 44 genes from intersection list.
Note: Top three enriched cellular processes with genes regulating these processes shown as nodes connected to cell processes by edges.

who went on to develop distant metastasis within five years.

Comparison with additional validation study

An additional breast cancer dataset[15] was used as a validation dataset to compare and confirm the results of the meta-analysis of the other three studies. Gene expression was compared between two groups of samples from this study. The first was a baseline group that included samples from ER+, node negative tumors without systemic treatment and without distant metastasis (DMFS time $>$ 5 years, 134 samples). The second group, the comparison group, included ER+, node negative tumors without systemic treatment with distant metastasis (DMFS time $<$ 5 years, 26 samples). A total of 685 genes were found to be differentially expressed (Additional file 8). This set of DEGs was further analyzed for enrichment of GO categories, as well as subnetwork analysis. The list of top enriched GO categories and cellular processes (Pathway Studio) was almost identical to the top enriched categories for the other three studies and for the IGS (Additional file 9). Comparison with the IGS of 44 genes showed significant overlap with 31 genes in common (Additional file 10) with a similar pattern

Table 4. Top 10 transcription factors ranked by significance of subnetwork enrichment among 44 genes.

Expression	Targets of neighbors	Total # of overlap	Percent overlap	Gene set seed	Overlapping entities	*P*-value
E2F4	68	6	8	E2F4	PLK1, CCNB1, BRCA1, CCNB2, MCM2, UBE2C	5.26E-08
FOXM1	133	7	5	FOXM1	PLK1, CCNB1, CDC25C, AURKA, CCNB2, NEK2, CENPA	1.26E-07
E2F1	243	8	3	E2F1	PTTG1, CCNB1, BRCA1, CDC25C, RACGAP1, MCM2, MELK, DUSP4	5.15E-07
BRCA1	83	5	5	BRCA1	PLK1, CCNB1, BRCA1, BUB1B, NEK2	5.04E-06
TP53	720	11	1	TP53	PTTG1, PLK1, CCNB1, BRCA1, CDC25C, HMMR, CCNB2, BUB1B, TOP2A, DUSP4, PRC1	5.40E-06
MYBL2	40	4	9	MYBL2	PLK1, CCNB1, TOP2A, UBE2C	6.75E-06
MASTL	2	2	66	MASTL	CCNB1, BUB1B	2.71E-05
CCAAT factors	199	6	3	CCAAT factors	PTTG1, PLK1, CCNB1, CDC25C, CCNB2, TOP2A	2.73E-05
E2F6	22	3	13	E2F6	CCNB1, BRCA1, CCNB2	4.41E-05
CDC5L	3	2	50	CDC5L	PLK1, CDC25C	5.41E-05

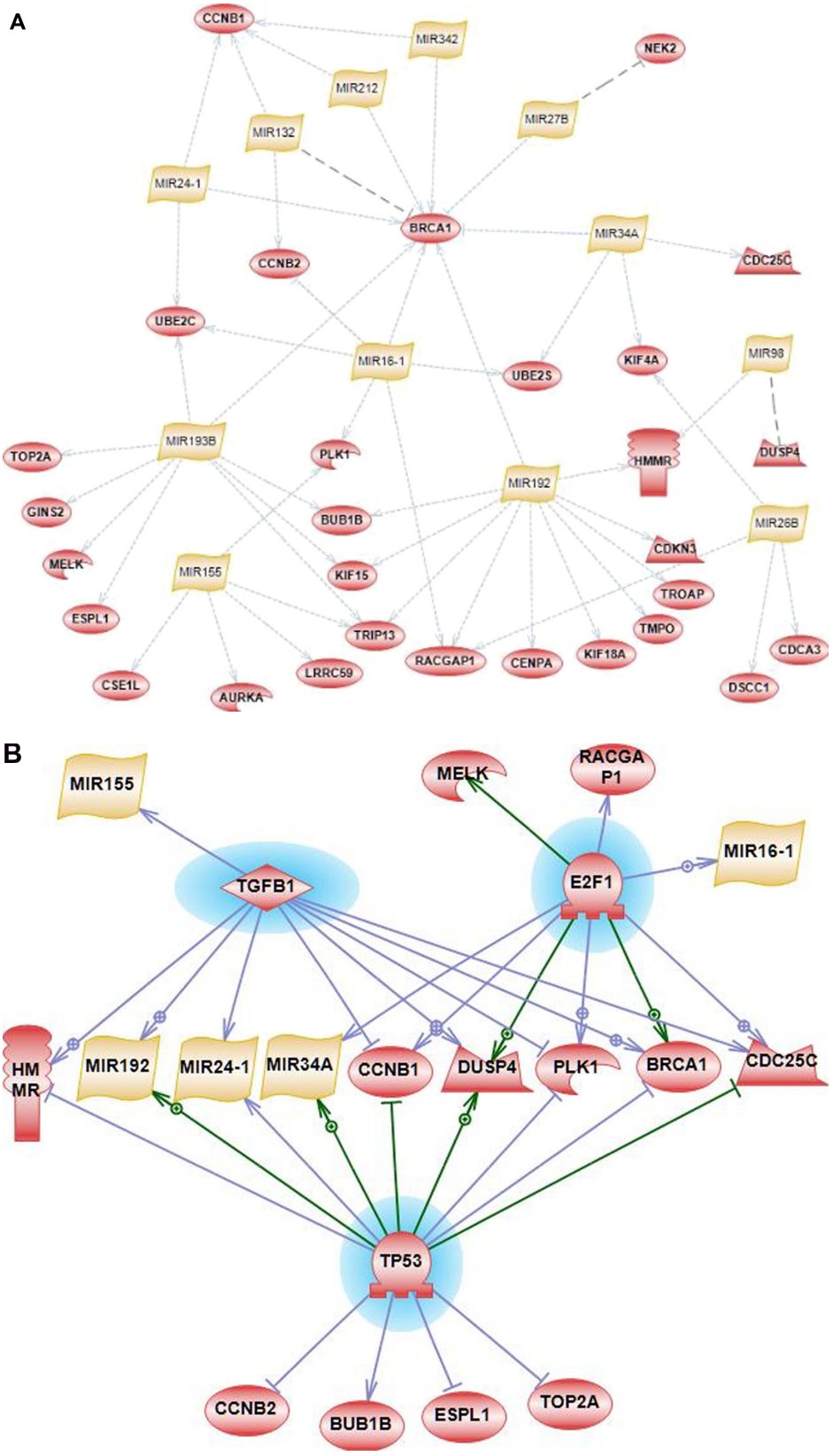

Figure 6. Subnetwork enrichment analysis of upstream transcriptional and post-transcriptional regulators. (**A**) 12 microRNAs (yellow nodes) were significantly enriched with targets (experimentally validated and/or predicted) among 44 genes (red nodes) from the intersection list. Only 3 interactions in this network are computationally predicted: miR-132 with BRCA1; miR-27B with NEK2; and miR-98 with DUSP4. These 3 predicted interactions are shown as dotted lines in grey color. All other interactions in this network are experientially validated (Tarbase 6.0) and shown as blue dotted lines. (**B**) Top three transcription factors regulating microRNA regulatory subnetwork (based on enrichment analysis of known transcription factors).

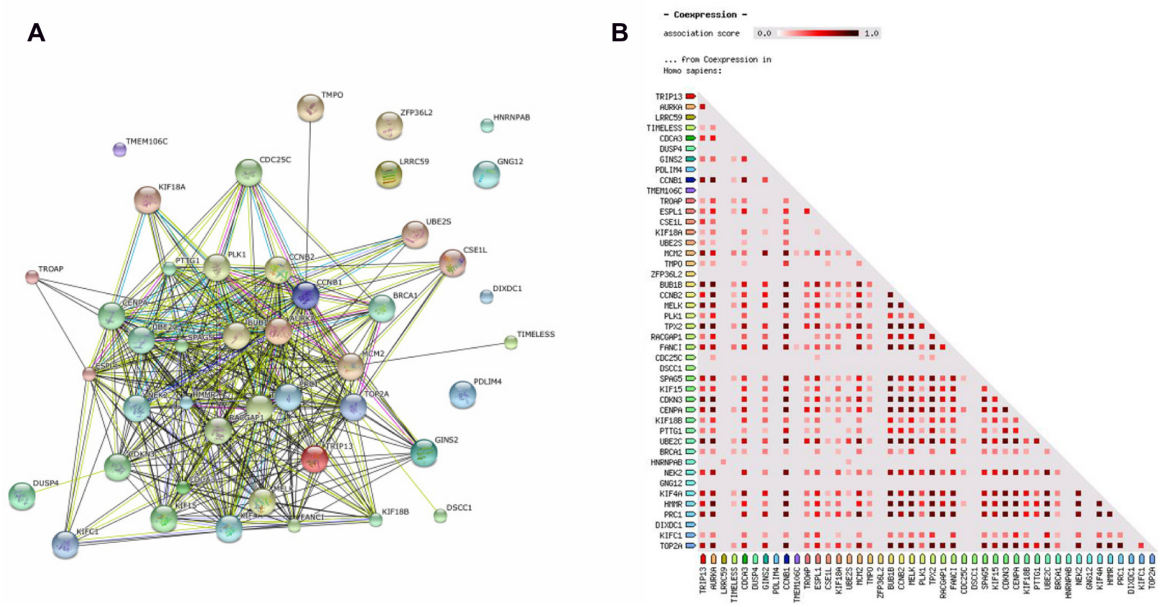

Figure 7. (**A**) STRING-generated protein-protein association networks within the 44 gene intersection list. (**B**) STRING-generated protein co-expression map (Homo sapiens) for 44 gene intersection list.

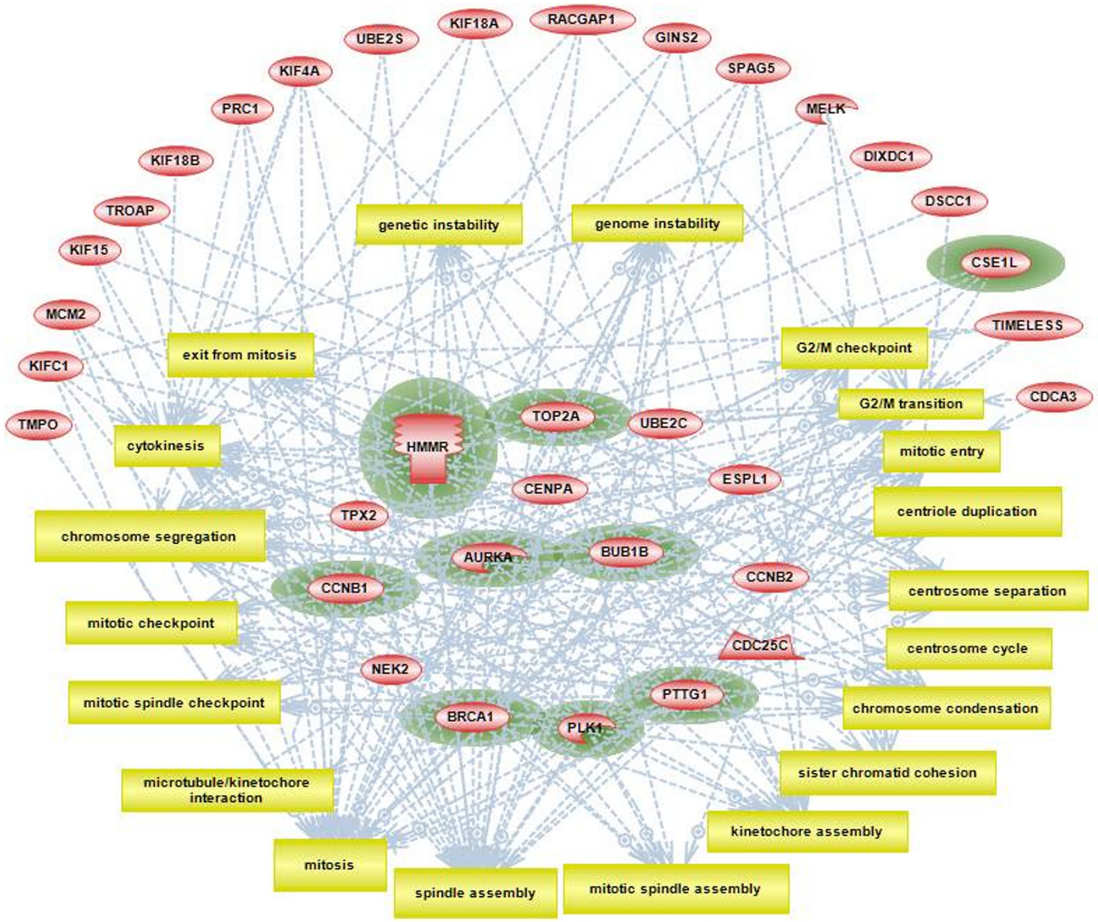

Figure 8. Pathway Studio analysis showing top 20 cellular processes all related to consecutive steps of progression through mitosis.
Notes: Nodes representing genes from the 44 intersection gene list are grouped according to number of downstream processes they are associated with: outside circle of nodes includes all nodes with <5 downstream processes; group of nodes in the middle includes highly interactive nodes with >5 processes. Genes that are significantly associated with metastasis phenotype (based on published data) are highlighted with green color.

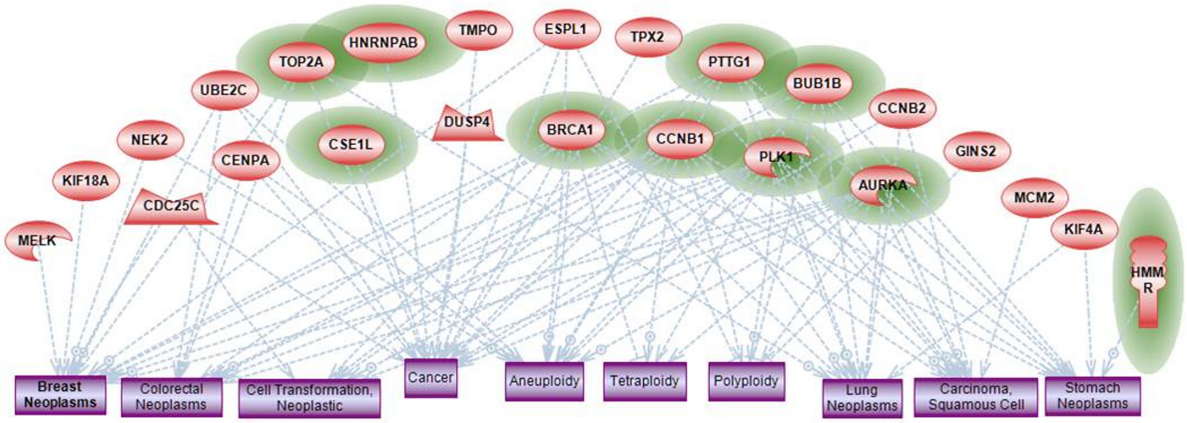

Figure 9. Enrichment analysis of disease/phenotype categories.
Notes: Groups of genes (red nodes) from 44 gene intersection list were identified that are overrepresented among proteins associated with various disease phenotypes. Top 10 significantly overrepresented disease categories are shown as rectangular nodes (blue color). Genes that are significantly associated with metastasis phenotype (based on published data) are highlighted with green color.

of expression—all but two genes were over expressed in the group with distant metastasis.

Discussion

The goal of this work was to apply bioinformatics approaches to publicly available gene expression data, with the end goal of identifying genes associated with early distant metastasis (<5 years) in ER+ breast cancer. Across three independent clinical studies, we found significant enrichment of multiple differentially expressed genes associated with many gene ontology categories, pathways, and biological processes that are associated with regulation of mitosis. A subset of 44 differentially expressed genes was common for all 3 studies. This intersection gene set is highly interactive and represents a significant part of the mitotic regulation network (MRN) of genes controlling all stages in the process of cell division. This process starts from G2/M entry checkpoint, chromosome condensation, centrosome and centriole duplication, mitotic spindle formation and organization, and includes exit from M-phase. Using GO enrichment analysis, we found over 20 different biological processes that were significantly enriched with genes from the MRN that were overexpressed in all three studies of breast cancer. However, we also observed that all top-ranked, enriched biological categories that are related to mitosis regulation were represented by additional genes that were not shared among the three studies. This is evident from the comparison of gene lists for each top functional category in all three studies (Fig. 1). While each GO category was represented and enriched by DEGs in all three studies, the number of genes varied between studies. This finding supports similar, recently reported observations for several types of cancer, where some of the aberrantly expressed genes sets were different between different samples whilst all mapping to a small number of the same biological pathways.[30] Therefore, such pathways and/or interacting gene networks could potentially serve as biomarkers of specific tumor phenotypes. In our meta-analysis we found the same top GO categories and pathways ranked by enrichment across each of three studies all related to regulation of mitotic check points and regulation.

The observation of common top ranked pathways across all three studies was reinforced by the analysis of an additional validation study.[15] All top GO categories were almost identical for the Schmidt study and each category was represented by some of the same (and also some different) genes from the same categories. While intersection with the 44 genes was significant (31 genes), additional functionally relevant genes were also found in many of the same GO groups. Similarly we found additional DEGs from the Schmidt study that were from the same top ranked pathways but were not common for all four studies. For example, four pathways involving Polo-like kinases were similarly significantly enriched in the Schmidt dataset, whereas some of the DEGs were different, albeit closely related, to the genes from the 44 IGS (Additional file 11).

The idea that mitotic checkpoint failure, deregulated mitosis, and/or impaired chromosome segregation is

indicative of poor outcome in many cancers, including those of the breast, is not new. For example, Hu et al recently reported that metastatic breast cancer, in both mouse and human, is characterized by a conserved gene signature that—specifically in ER+ tumors—is linked to the mitotic spindle checkpoint.[31] This has implications for anti-mitotic agents and spindle poisons such as vinca alkaloids and taxanes, which have yet to be proven to be 100% effective in preventing tumor formation.[32] It has been shown that spindle poisons prevent cancer cells from forming spindle microtubules and cause them to go into apoptosis.[33]

It has long been known that mitosis-regulating genes promote normal cell division and ensure chromosomal and genomic stability. It was also believed that loss of function mutations or deletions are required to impair mitotic checkpoint regulation.[34,35] However, studies have indicated that very few mutations are found among mitotic checkpoint genes in most human tumors.[36–39] The most commonly reported pattern was that of overexpression in tumors, including many of the genes we report here.[40,41] Much evidence indicates that overexpression of these checkpoint genes in fact leads to impairment of regulatory controls and increased mitotic activity of tumor cells.[41]

Several of the highly interactive nodes of the mitotic regulation network reported here are overexpressed in breast cancer and implicated in generation of chromosomal and genomic instability. Overexpression of the Aurora-A kinase in mammalian cells leads to centrosome amplification, genetic instability, and transformation.[42] Inhibition of Aurora-A has been shown to lead to increased cisplatin sensitivity.[43] Similar observations have been reported on human PTTG1 (hPTTG1), the index mammalian securin. PTTG1 is a critical component of the spindle checkpoint controlling faithful chromatid separation and has also been identified as a proto-oncogene.[44] hPTTG1 abundance[45] or loss of function,[46] both result in abnormal mitosis and chromosomal instability. The mitotic-spindle/hyaluronan-binding protein RHAMM, an essential component of normal mitotic machinery, is overexpressed in many human tumors. Intracellular RHAMM associates with BRCA1 and BARD1; this association attenuates the mitotic-spindle-promoting activity of RHAMM that might contribute to tumor progression by promoting genomic instability.[47]

Another important player, Polo-like kinase (PLK1), was consistently overexpressed in each of the three studies in our comparison and is known to be a key regulator of mitotic pathways such as centrosome cycle control (Fig. 3). Overexpression of PLK1 is shown to be associated with increased chromosomal instability, while its inhibition prevents the growth of metastatic breast cancer cells in the brain.[48]

These findings also suggest a connection to the metastatic phenotype of breast tumor cells. Overall, our enrichment analysis has shown that 10 of 44 genes that were overexpressed in all three studies are experimentally associated with the metastasis phenotype; it was also shown that nine of these genes are major regulatory hubs of the mitosis regulatory network reported. Importantly, dysregulation of these hubs has been connected to abnormal mitosis and genomic instability.

Missegregation of chromosomes is a hallmark of genomic instability,[49] and chromosomal instability (CIN) as measured by single nucleotide polymorphism analysis of breast tumors is significantly associated with time to distant metastasis in ER+ disease.[50] However, our discovery of a common network of highly interconnected genes controlling the entire sequence of mitotic events, consistently overexpressed in all three studies, provides insight into potential large scale deregulation of the mitotic machinery that is associated with, and might even facilitate, a progression toward early metastasis.

Analysis of common upstream regulators of the mitotic regulatory network

We also identified several points of transcriptional and post-transcriptional control of this mitotic regulation network.

Transcriptional analysis

The promoter regions of more than 30% of the genes in the IGS are enriched for NFY transcription factor binding sites. Six out of 44 have been reported in literature as experimentally validated targets of NFYA, NFYB, or NFYC (Additional file 7). An association between breast cancer progression[51] or metastasis[52] and enrichment of binding sites for NFY has been previously reported. NFY can cooperate with ER (via SP1 sites) to regulate transcription of E2F1, which is also enriched in our DEG list.

However, estrogen-stimulated activation of NFY requires non-genomic effects of ER, specifically through cAMP and PKA.[53]

Interestingly, there is documented evidence that connects poor response to Tamoxifen (the only systemic therapy received by some patients in this study) and non-genomic (ie, non-transcriptional), ER-dependent signaling, which often involves kinases such as Src, Mapk, and Akt (reviewed in literature).[54]

E2F sites were found to be significantly enriched in the promoter regions of the 44 genes in the IGS using MSigDB analysis, which is in concordance with the TF and subnetwork analyses using Ingenuity, and Pathway Studio, respectively. The main characterized function of the E2F family of transcription factors is the regulation of the cell cycle, as well as genes involved in DNA synthesis, repair, apoptosis, development, and differentiation. It is therefore understandable that E2F sites are enriched in the IGS of the three studies. Several members of the E2F family regulate expression of MYBL2, a transcription factor on our list that regulates the expression of genes involved in cancer progression. E2F4—which was the top ranked TF we found in our study (Table 4)—has primarily been described as a repressor of both transcription and cell cycle progression,[55–57] but it can also be a transcriptional activator. Overexpression of E2F4 in colorectal cancer cells has been shown to contribute to hyperproliferation.[58–60] A more recent study suggested that low E2F4 gene expression and is predictive for survival of ER negative breast cancer patients.[61] E2F transcription factor1 (E2F1), the third top ranked TF from our study, is responsible for the regulation of FOXM1 expression (the second top ranked TF here), which plays a key role in epirubicin resistance in cancer cells.[62] Another study has shown that E2F1 expression was enhanced in Tamoxifen-resistant MCF-7 breast cancer cells.[63]

The FOXM1 transcription factor is a transcriptional activator involved in several biological processes, including cell proliferation and differentiation, cell cycle progression, DNA repair, angiogenesis, and apoptosis. Elevated FOXM1 expression is found in numerous cancers and studies have found FOXM1 to be one of the most commonly upregulated genes in human solid tumors,[64] thus it is not surprising this was the second top ranked TF found in our analysis. Third and fourth on the list were Breast Cancer Type 1 susceptibility protein (BRCA1) and Tumor Protein p53 (TP53). BRCA1 is involved in DNA repair and tumor suppression through interaction with a large network of DNA repair proteins. Mutations in BRCA1 are found in 5%–50% of familial breast cancers and are often associated with increased risk. Studies suggest that TP53 mutations may precede loss of the BRCA1 wild-type allele.[65] TP53 mutations are often found in BRCA1 mutant tumors and play a large role in tumor suppression. Mutant p53 genes are found in a number of cancers and cause the loss of tumor suppressor activity. Another TF on our list—cell division cycle 5-like protein (Cdc5L)—plays a key role in regulation of cell division. It is also thought to suppress p53-induced growth arrest.[66]

Several CCAAT factors were seen as top hits across the three studies (Table 4). CCAAT-enhancer binding proteins are transcriptional activators involved in numerous biological processes, including cellular proliferation. A recent study has shown that CCAAT/enhancer binding protein delta CpG island methylation is associated with metastasis in breast cancer.[67] Finally, one of the TFs on the list that is not a well know regulator is microtubule-associated serine/threonine kinase-like (MASTL), which is the human ortholog of the Greatwall kinase that has been shown to be most active in mitosis, facilitating mitotic entry, anaphase, and cytokinesis.[68]

microRNA analysis

Post-transcriptional gene regulation by small, non-coding RNAs (microRNAs) are also of considerable interest due to their significance in numerous biological processes. Currently studies indicate that the primary mechanism of microRNA interaction with their mRNA targets is through translational inhibition, although there is some debate about the role of microRNAs affecting mRNA stability and promoting degradation.[69] Twelve distinct microRNAs were found to have their target sets significantly enriched with 29 of the 44 genes, with miR-192 targets being the most over-represented. All but three of these microRNA to target interactions have been experimentally validated and previously reported (Tarbase 6.0). Several of these microRNAs and their targets are well characterized as being associated

with breast cancer. Specifically, miR-192 has been shown to be one of the p53 inducible microRNAs that controls multiple cell cycle checkpoints and inhibits cell proliferation when overexpressed.[70,71] In addition to cell cycle regulation, it is interesting to note that p53-mediated upregulation of miR-192 prevents the epithelial-mesenchymal transition (an in vitro corollary of metastatic potential) in hepatocellular carcinoma through repression of the transcription factor ZEB2.[72] Another important microRNA regulatory network was detected among the 44 genes that includes miR-193B and its experimentally validated targets (the second highest ranked microRNA with eight genes targeted). It has been recently shown that miR-193B targets estrogen receptor-α (ERα) and inhibits estrogen-induced growth of breast cancer cells.[73–75] Importantly, miR-193B is a known moderator of ER signaling. Also of note, reduced expression of miR-342 has been associated with tamoxifen-resistant breast tumors and breast cancer cell line MCF-7/HER2Δ16 and MCF-7 variants TAMR1 and LCC2.[76] Two other microRNAs from this set, miR-212 and miR-132 (miR212/miR132 family), were shown to regulate epithelial-stromal interactions during development of the mammary gland.[77] They were also reported as over-expressed in pancreatic adenocarcinoma and targeting retinoblastoma tumor suppressor Rb1.[78]

Therapeutic targets

Protein products of three of the 44 genes are targets of specific therapeutic agents; each of these three is also implicated as prognostic or predictive biomarkers in ER+ breast cancer. Topoisomerase II alpha (TOPO2A, more commonly referred to as TOP2A) is a target of the anthracycline class of chemotherapeutic agents (doxorubicin/adriamycin, epirubicin, etc.) and a poor prognostic factor in ER+ breast cancer that is associated with a proliferative index (Ki67 labeling).[79] However, the strength of TOPO2A amplification or overexpression as a predictive factor for beneficial response to anthracyclines remains unclear.[80] PLK1 is more commonly expressed in ER– breast tumors[81] and was recently identified in an siRNA library screen as a selective inhibitor of tumor-initiating cells in triple negative breast cancer cell lines.[26] However, in women with ER+ breast cancer, upregulation of a 14-3-3ζ/YWHAZ-centered network predicts poor response to Tamoxifen and distant metastasis; PLK1 is a major component of this network.[82] Finally, AURKA is a key component of SCMGENE, a 3-gene signature (ER, HER2, AURKA) that performs as well, or better, than the original gene signatures developed to classify basal, HER2+, luminal A, and luminal B breast cancer.[83] Down regulation of AURKA can also reverse estrogen-mediated growth in breast cancer cells.[84]

Conclusion

A meta-analysis of three independent ER+ breast cancer studies has shown a common pattern of expression for 44 genes involved with the regulation of most stages of mitosis, starting from G2/M transition, mitotic entry, specific processes of chromosome condensation, centrosome and centriole duplication, spindle assembly, chromosome attachment, alignment, and segregation. The pattern of expression of these common genes is remarkably similar in all three datasets, with almost all genes being overexpressed in patients with early metastasis. This cluster of genes encodes proteins that are highly interactive and known to be involved with multiple checkpoints regulating entrance and transition through major phases of mitosis. Extensive literature mining has shown that overexpression of a significant number of these proteins is associated with abnormal mitotic events that lead to abnormal cell division and consequently an increased rate of chromosomal and genomic instability. Importantly, our findings indicate that this gene cluster is differentially expressed in patients with early metastasis; therefore, we hypothesize that correlated overexpression of multiple key regulators of mitotic checkpoints could be indicative of early onset and increase of genomic instability and a progression towards a metastatic cell phenotype. Ten of the genes common to all three studies were already reported as associated with a metastatic phenotype. Few details are known with regard to specific mechanisms of how genomic instability contributes to such progression. However, our meta-analysis has narrowed down a list of potential players and their transcriptional and post-transcriptional regulators, thus providing additional evidence for further mechanistic studies and validation experiments.

Acknowledgements
We thank Dr. Juli Klemm (NCI), Dr. Kevin Groch (SAIC-F), and Dr. Debra Hope (SAIC-F) for supporting the In Silico Research Centers of Excellence (ISRCE) program.

Funding
This study was funded primarily by HHSN2612200800001E, the caBIG In Silico Research Centers of Excellence (ISRCE) program from NCI/SAIC. Work was also funded in part by Susan G. Komen for the Cure (KG090187 to RBR).

Author Contributions
Conceived and designed the experiments: YG, RBR, RC, SM. Analyzed data: YG, RBR, KB, RG. Wrote the first draft of the manuscript: YG, RBR, SM. Contributed to the writing of the manuscript: YG, RBR, SM, LS, RC. Agree with manuscript results and conclusions: YG, RBR, RG, KB, LS, RC, SM. Jointly developed the structure and arguments for the paper: YG, RBR, SM. Made critical revisions and approved final version: YG, RBR, SM, LS. All authors reviewed and approved of the final manuscript.

Competing Interests
Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics
As a requirement of publication, the author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References
1. American Cancer Society. Cancer Facts & Figures 2012. Atlanta: American Cancer Society; 2012.
2. Thorpe SM. Estrogen and progesterone receptor determinations in breast cancer. Technology, biology and clinical significance. *Acta Oncol*. 1988;27(1):1–19.
3. Thompson A, Brennan K, Cox A, et al. Evaluation of the current knowledge limitations in breast cancer research: a gap analysis. *Breast Cancer Res*. 2008;10(2):R26.
4. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*. 2008;452(7187): 564–70.
5. Chanrion M, Negre V, Fontaine H, et al. A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clin Cancer Res*. 2008;14(6):1744–52.
6. Goetz MP, Suman VJ, Ingle JN, et al. A two-gene expression ratio of homeobox 13 and interleukin-17B receptor for prediction of recurrence and survival in women receiving adjuvant tamoxifen. Clin Cancer Res. 2006; 12(7 Pt 1):2080–7.
7. Jansen MP, Foekens JA, van Staveren IL, et al. Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. *J Clin Oncol*. 2005;23(4):732–40.
8. 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.
9. Reid JF, Lusa L, De Cecco L, et al. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst*. 2005;97(12):927–30.
10. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst*. 2010;102(7): 464–74.
11. Madhavan S, Gusev Y, Harris M, et al. G-DOC: a systems medicine platform for personalized oncology. *Neoplasia*. 2011;13(9):771–83.
12. Demicheli R, Ardoino I, Boracchi P, et al. Recurrence and mortality according to estrogen receptor status for breast cancer patients undergoing conservative surgery. Ipsilateral breast tumour recurrence dynamics provides clues for tumour biology within the residual breast. *BMC cancer*. 2010;10:656.
13. 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.
14. Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*. 2007;13(11):3207–14.
15. Schmidt M, Bohm D, von Torne C, et al. The humoral immune system has a key prognosticimpact in node-negative breast cancer. *Cancer Res*. 2008;68(13):5405–13.
16. 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.
17. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.
18. Wettenhall JM, Smyth GK. LIMMAGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*. 2004;20(18):3705–6.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*. 1995;57:289–300.
20. Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011;39(Database issue): D691–7.
21. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*. 2011;39(Database issue):D691–7.
22. Chang J, Clark GM, Allred DC, Mohsin S, Chamness G, Elledge RM. Survival of patients with metastatic breast carcinoma: importance of prognostic markers of the primary tumor. *Cancer*. 2003;97(3):545–53.
23. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–40.

24. Vergoulis T, Vlachos IS, Alexiou P, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*. 2012;40(Database issue):D222–9.
25. Fountzilas G, Valavanis C, Kotoula V, et al. HER2 and TOP2A in high-risk early breast cancer patients treated with adjuvant epirubicin-based dose-dense sequential chemotherapy. *J Trans Med*. 2012;10:10.
26. Hu K, Law J, Fotovati A, Dunn SE. Small interfering RNA library screen identified polo-like kinase-1 (PLK1) as a potential therapeutic target for breast cancer that uniquely eliminates tumour-initiating cells. *Breast Cancer Res*. 2012;14(1):R22.
27. Schoffski P, Blay JY, De Greve J, et al. Multicentric parallel phase II trial of the polo-like kinase 1 inhibitor BI 2536 in patients with advanced head and neck cancer, breast cancer, ovarian cancer, soft tissue sarcoma and melanoma. The first protocol of the European Organization for Research and Treatment of Cancer (EORTC) Network Of Core Institutes (NOCI). *Eur J Cancer*. 2010;46(12):2206–15.
28. Macarulla T, Cervantes A, Elez E, et al. Phase I study of the selective Aurora A kinase inhibitor MLN8054 in patients with advanced solid tumors: safety, pharmacokinetics, and pharmacodynamics. *Mol Cancer Ther*. 2010;9(10):2844–52.
29. Sehdev V, Peng D, Soutto M, et al. The Aurora Kinase A inhibitor MLN8237 enhances cisplatin-induced cell death in esophageal adenocarcinoma cells. *Mol Cancer Ther*. 2012;11(3):763–74.
30. Leary RJ, Lin JC, Cummins J, et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A*. 2008;105(42):16224–9.
31. Hu Y, Wu G, Rusch M, et al. Integrated cross-species transcriptional network analysis of metastatic susceptibility. *Proc Natl Acad Sci U S A*. 2012;109(8):3184–9.
32. Wood KW, Cornwell WD, Jackson JR. Past and future of the mitotic spindle as an oncology target. *Curr Opin Pharmacol*. 2001;1(4):370–7.
33. Matson DR, Stukenberg PT. Spindle poisons and cell fate: a tale of two pathways. *Mol Interv*. 2011;11(2):141–50.
34. Bharadwaj R, Yu H. The spindle checkpoint, aneuploidy, and cancer. *Oncogene*. 2004;23(11):2016–27.
35. Cahill DP, Lengauer C, Yu J, et al. Mutations of mitotic checkpoint genes in human cancers. *Nature*. 1998;392(6673):300–3.
36. Gemma A, Hosoya Y, Seike M, et al. Genomic structure of the human MAD2 gene and mutation analysis in human lung and breast cancers. *Lung Cancer*. 2001;32(3):289–95.
37. Ingvarsson S, Sigbjornsdottir BI, Huiping C, et al. Mutation analysis of the CHK2 gene in breast carcinoma and other cancers. *Breast Cancer Res*. 2002;4(3):R4.
38. Miller CW, Ikezoe T, Krug U, et al. Mutations of the CHK2 gene are found in some osteosarcomas, but are rare in breast, lung, and ovarian tumors. *Gene Chromosomes Cancer*. 2002;33(1):17–21.
39. Myrie KA, Percy MJ, Azim JN, Neeley CK, Petty EM. Mutation and expression analysis of human BUB1 and BUB1B in aneuploid breast cancer cell lines. *Cancer Lett*. 2000;152(2):193–9.
40. Grabsch H, Takeno S, Parsons WJ, et al. Overexpression of the mitotic checkpoint genes BUB1, BUBR1, and BUB3 in gastric cancer—association with tumour cell proliferation. *J Pathol*. 2003;200(1):16–22.
41. Yuan B, Xu Y, Woo JH, et al. Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability. *Clin Cancer Res*. 2006;12(2):405–10.
42. Lentini L, Amato A, Schillaci T, Di Leonardo A. Simultaneous Aurora-A/STK15 overexpression and centrosome amplification induce chromosomal instability in tumour cells with a MIN phenotype. *BMC cancer*. 2007;7:212.
43. Katayama H, Sasai K, Kawai H, et al. Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53. *Nat Genet*. 2004;36(1):55–62.
44. Zou H, McGarry TJ, Bernal T, Kirschner MW. Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science*. 1999;285(5426):418–22.
45. Yu R, Lu W, Chen J, McCabe CJ, Melmed S. Overexpressed pituitary tumor-transforming gene causes aneuploidy in live human cells. *Endocrinology*. 2003;144(11):4991–8.
46. Wang Z, Moro E, Kovacs K, Yu R, Melmed S. Pituitary tumor transforming gene-null male mice exhibit impaired pancreatic beta cell proliferation and diabetes. *Proc Natl Acad Sci U S A*. 2003;100(6):3428–32.
47. Maxwell CA, McCarthy J, Turley E. Cell-surface and mitotic-spindle RHAMM: moonlighting or dual oncogenic functions? *J Cell Sci*. 2008;121(Pt 7):925–32.
48. Qian Y, Hua E, Bisht K, et al. Inhibition of Polo-like kinase 1 prevents the growth of metastatic breast cancer cells in the brain. *Clin Exp Metastasis*. 2011;28(8):899–908.
49. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet*. 2012;13(3):189–203.
50. Smid M, Hoes M, Sieuwerts AM, et al. Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. *Breast Cancer Res Treat*. 2011;128(1):23–30.
51. Niida A, Smith AD, Imoto S, et al. Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells. *BMC Bioinformatics*. 2008;9:404.
52. Thomassen M, Tan Q, Kruse TA. Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer. *BMC Cancer*. 2008;8:394.
53. Ngwenya S, Safe S. Cell context-dependent differences in the induction of E2F-1 gene expression by 17 beta-estradiol in MCF-7 and ZR-75 cells. *Endocrinology*. 2003;144(5):1675–85.
54. Riggins RB, Schrecengost RS, Guerrero MS, Bouton AH. Pathways to tamoxifen resistance. *Cancer Lett*. 2007;256(1):1–24.
55. Vairo G, Livingston DM, Ginsberg D. Functional interaction between E2F-4 and p130: evidence for distinct mechanisms underlying growth suppression by different retinoblastoma protein family members. *Genes Dev*. 1995;9(7):869–81.
56. Muller H, Moroni MC, Vigo E, Petersen BO, Bartek J, Helin K. Induction of S-phase entry by E2F transcription factors depends on their nuclear localization. *Mol Cell Biol*. 1997;17(9):5508–20.
57. Rayman JB, Takahashi Y, Indjeian VB, et al. E2F mediates cell cycle-dependent transcriptional repression in vivo by recruitment of an HDAC1/mSin3B corepressor complex. *Genes Dev*. 15 2002;16(8):933–47.
58. Mady HH, Hasso S, Melhem MF. Expression of E2F-4 gene in colorectal adenocarcinoma and corresponding covering mucosa: an immunohistochemistry, image analysis, and immunoblot study. *Appl Immunohistochem Mol Morphol*. 2002;10(3):225–30.
59. Garneau H, Alvarez L, Paquin MC, et al. Nuclear expression of E2F4 induces cell death via multiple pathways in normal human intestinal epithelial crypt cells but not in colon cancer cells. *Am J Physiol Gastrointest Liver Physiol*. 2007;293(4):G758–72.
60. Garneau H, Paquin MC, Carrier JC, Rivard N. E2F4 expression is required for cell cycle progression of normal intestinal crypt cells and colorectal cancer cells. *J Cell Physiol*. 2009;221(2):350–8.
61. Koch M, Hanl M, Wiese M. Feature extraction via composite scoring and voting in breast cancer. *Breast Cancer Res Treat*. 2012;135(1):307–18.
62. de Olano N, Koo CY, Monteiro LJ, et al. The p38 MAPK-MK2 Axis Regulates E2F1 and FOXM1 Expression after Epirubicin Treatment. *Mol Cancer Res*. 2012;10(9):1189–202.
63. Lee KY, Lee JW, Nam HJ, Shim JH, Song Y, Kang KW. PI3-kinase/p38 kinase-dependent E2F1 activation is critical for Pin1 induction in tamoxifen-resistant breast cancer cells. *Mol Cells*. 2011;32(1):107–11.
64. Koo CY, Muir KW, Lam EW. FOXM1: From cancer initiation to progression and treatment. *Biochim Biophys Acta*. 2012;1819(1):28–37.
65. Jonkers J. Tracking evolution of BRCA1-associated breast cancer. *Cancer Discov*. 2012;2(6):486–8.
66. Zhang N, Kaur R, Akhter S, Legerski RJ. Cdc5L interacts with ATR and is required for the S-phase cell-cycle checkpoint. *EMBO reports*. 2009;10(9):1029–35.
67. Palmieri C, Monteverde M, Lattanzio L, et al. Site-specific CpG methylation in the CCAAT/enhancer binding protein delta (CEBPdelta) CpG island in breast cancer is associated with metastatic relapse. *Br J Cancer*. 2012;107(4):732–8.

68. Voets E, Wolthuis RM. MASTL is the human orthologue of Greatwall kinase that facilitates mitotic entry, anaphase and cytokinesis. *Cell Cycle*. 2010;9(17):3591–601.
69. Djuranovic S, Nahvi A, Green R. A parsimonious model for gene regulation by miRNAs. *Science*. 2011;331(6017):550–3.
70. Braun CJ, Zhang X, Savelyeva I, et al. p53-Responsive micrornas 192 and 215 are capable of inducing cell cycle arrest. *Cancer Res*. 2008;68(24): 10094–104.
71. Georges SA, Biery MC, Kim SY, et al. Coordinated regulation of cell cycle transcripts by p53-Inducible microRNAs, miR-192 and miR-215. *Cancer Res*. 2008;68(24):10105–12.
72. Kim T, Veronese A, Pichiorri F, et al. p53 regulates epithelial-mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. *J Exp Med*. 2011;208(5):875–83.
73. Leivonen SK, Makela R, Ostling P, et al. Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*. 2009;28(44):3926–36.
74. Leivonen SK, Rokka A, Ostling P, et al. Identification of miR-193b targets in breast cancer cells and systems biological analysis of their functional impact. *Mol Cell Proteomics*. 2011;10(7):M110.005322.
75. Yoshimoto N, Toyama T, Takahashi S, et al. Distinct expressions of microRNAs that directly target estrogen receptor alpha in human breast cancer. *Breast Cancer Rest Treat*. 2011;130(1):331–9.
76. Cittelly DM, Das PM, Spoelstra NS, et al. Downregulation of miR-342 is associated with tamoxifen resistant breast tumors. *Mol Cancer*. 2010;9:317.
77. Ucar A, Vafaizadeh V, Jarry H, et al. miR-212 and miR-132 are required for epithelial stromal interactions necessary for mouse mammary gland development. *Nat Genet*. 2010;42(12):1101–8.
78. Park JK, Henry JC, Jiang J, et al. miR-132 and miR-212 are increased in pancreatic cancer and target the retinoblastoma tumor suppressor. *Biochem Biophys Res Commun*. 2011;406(4):518–23.
79. Tokiniwa H, Horiguchi J, Takata D, et al. Topoisomerase II alpha expression and the Ki-67 labeling index correlate with prognostic factors in estrogen receptor-positive and human epidermal growth factor type-2-negative breast cancer. *Breast Cancer*. 2012;19(4):309–14.
80. Romero A, Caldes T, Diaz-Rubio E, Martin M. Topoisomerase 2 alpha: a real predictor of anthracycline efficacy? *Clin Transl Oncol*. 2012;14(3):163–8.
81. Weichert W, Kristiansen G, Winzer KJ, et al. Polo-like kinase isoforms in breast cancer: expression patterns and prognostic implications. *Virchows Arch*. 2005;446(4):442–50.
82. Bergamaschi A, Christensen BL, Katzenellenbogen BS. Reversal of endocrine resistance in breast cancer: interrelationships among 14-3-3zeta, FOXM1, and a gene signature associated with mitosis. *Breast Cancer Res*. 2011;13(3):R70.
83. Haibe-Kains B, Desmedt C, Loi S, et al. A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst*. 2012;104(4): 311–25.
84. Lee HH, Zhu Y, Govindasamy KM, Gopalan G. Downregulation of Aurora-A overrides estrogen-mediated growth and chemoresistance in breast cancer cells. *Endocr Relat Cancer*. 2008;15(3):765–75.

Supplementary Data

Additional file 1 (PDF)

Description: Figure showing distant metastasis (DM)-free survival proportions for three publicly available breast cancer datasets. Selected patients (ER+, with either documented distant metastasis $\leq$ 5 years or no distant metastasis $>$ 5 years) from each study were analyzed by the Kaplan-Meier estimator. $P = 0.32$, not significant, by log rank test.

Additional file 2 (PDF)

Description: Figure showing primary tumor size is significantly larger in Group 2 (distant metastasis-positive patients) than in Group 1 (distant metastasis-negative patients) for the Desmedt (A), Sotiriou (B) and Loi (C) datasets. $*P \leq 0.05$ and $***P \leq 0.0001$ by Mann-Whitney rank sum test.

Additional file 3 (Excel)

Description: Tables with results of differential gene expression analysis for each of 3 studies. Fold change represents Group 2 (distant metastasis at $\leq$5 years) vs. Group 1 (no documented metastasis after 5 years).

Additional file 4 (Excel)

Description: Intersection gene list with annotations.

Additional file 5 (Excel)

Description: Subnetwork enrichment analysis results of cell processes downstream of the 44 intersection genes. (Pathway Studio).

Additional file 6 (Excel)

Description: Reactome-generated pathway enrichment results for the 44 intersection genes.

Additional file 7 (PDF)

Description: NFY regulatory network identified by Ingenuity.

Additional file 8 (Excel)

Description: Table with results of differential gene expression analysis for additional validation study Schmidt et al. Fold change represents Group 2 (distant metastasis at $\leq$5 years) vs. Group 1 (no documented metastasis after 5 years).

Additional file 9 (Excel)

Description: Subnetwork enrichment analysis results of biological processes downstream of the differentially expressed genes from additional validation study Schmidt et al. (Pathway Studio).

Additional file 10 (Excel)

Description: Intersection gene list between set of DEGs from additional study Schmidt et al and 44 genes from IGS with annotations.

Additional file 11 (PDF)

Description: Results of pathway enrichment analysis (Ingenuity Pathway Analysis 8.5) for DEGs from the additional validation study Schmidt et al. One of the top ranked significantly enriched group shows four pathways related to Polo-like Kinase regulation of mitotic events. Differentially expressed genes are shown in grey color.