Towards pathway and network-based medicine in breast cancer

Himanshu Joshi

Faculty of Medicine
Institute of Clinical Medicine
University of Oslo

Oslo

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1734

ISBN 978-82-8264-820-2

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Abstract

Introduction

Breast cancer is a molecularly heterogeneous disease. The existing molecular classifications provide a good introduction of the molecular heterogeneity. However more efforts are necessary to characterize molecularly distinct tumor subgroups that are likely responders to novel targeted therapy interventions. One of the classification strategies is to characterize robust classes based on activity of specific biological pathways and networks. This thesis describes a classification based on activity status of p53, ER and VEGF signaling in breast cancer.

Methods

In paper I, canonical sequences from the proximal promoter of the genes that distinguish the molecular portraits are studied to identify the significantly overrepresented potential TFBS motifs for each molecular class. Subtype-specific networks based on previously reported or predicted interactions between subtype-specific genes and corresponding transcription factors. In paper II, breast cancer expression profiles were analyzed for inferring differentially activated pathways and differentially expressed genes by p53 gene mutation status using geneset-based and individual gene search methods. Genes are evaluated for prognostic significance. Paper III compares miRNA and mRNA expression profiles from the same sample set by VEGF mRNA expression status and for each differentially expressed miRNA, class-specific potential targets are the mRNAs having target site (based on the target database) and are differentially expressed as well as class-specific anti-correlated to their potential regulator miRNA.

Results

We identified the significantly overrepresented potential TFBS motifs by subtype and showed positive correlation between the subtype-specific mRNA expressions of some of their corresponding TF genes and degree of TFBS overrepresentation. The network analysis showed p53 as a topological hub that has interactions with subtype-specific genes thus explaining core functional significance of p53 signaling (Paper I). We also identified about 40 pathways differentially activated by the *p53* mutation status. Besides *VEGF* expression was shown to predict survival after controlling for *p53* mutation status and subtype. In ER+/PR+ patients, effect of *VEGF* was found significant but not in ER–/PR– patients (Paper II). MiRNA profiles of VEGF upregulated group showed upregulation of miR-590-5p, miR-18a/18b/19a cluster, miR-9/9*, miR-135b, and downregulation of miR-149, miR-342-3p/5p, miR- 449a. The anti-correlated targets of upregulated miRNAs were enriched for angiogenesis pathway, vasculature development, TGF-β signaling and focal adhesion. Anti-correlated targets of downregulated miRNAs in VEGFA+ group were associated with EGFR pathway, positive regulation of DNA binding and nucleolus (Paper III). This work implicates experimental validation.

Keywords: Breast cancer, p53 signaling, molecular classification, VEGF signaling, pathway analysis, transcriptional regulation, miRNAs, miRNA regulatory modules

Acknowledgements

I would like to thank my PhD advisor, Director of Statistics for Innovation and Professor of statistics, Arnoldo Frigessi, for providing me his support when it was needed. I would also acknowledge his help by gainful discussions on biostatistical aspects. Without his guidance and direction this work would not have come to realization. His guidance also helped me grow confident as an independent researcher.

I like to thank the PhD committee for their useful suggestions.

I like to thank Prof Eivind Hovig, who took interest in my work and provided help whenever necessary.

I also thank all professors, co-advisors and the staff, especially Karin and Nina at Research Center (Akershus University Hospital) for providing me all necessary support during my PhD program.

I am thankful to all library (Ahus) staff, especially Senior Librarian Helene Lie for granting me an access to relevant databases and helping me with effective literature search.

I also like to thank Professor Emeritus Pieter Kierulf for offering me his advice whenever needed.

I like to thank the Regional Health Authority of South-Eastern Norway for providing the research funding.

I like to thank my parents for being a constant source of inspiration during all this time.

I also like to thank my sister, who provided me her support and valuable scientific input on my work.

Himanshu Joshi

Preface

Significant technological advances in the past decade have enhanced the potential to study the molecularly heterogeneous diseases such as breast cancer. The goals of breast cancer management are: to increase the rates of pathologic complete response, prevention of metastasis, relapse and recurrence, increase the relapse free and overall survival of patients and minimize the treatment-associated adverse effects. In the current scenario, both chemotherapy and available targeted therapy options (for instance Tamoxifen, Herceptin) are associated with significant proportion of failure or resistance. The possible explanation for the treatment failure is lack of consideration to pathways, networks and their feedbacks that are responsible for shaping the overall phenotype. Efficient individualized cancer management requires the focus on three main goals: 1. Characterization of tumor classification based on the combination of molecular alterations in each individual patient and then formulating the cancer management strategy tailored to the individual's genetic profile and tumor characteristics. 2. Individualizing the treatment choices and drug doses to minimize the treatment associated adverse effects and chances of metastasis and thereby improving patients' quality of life. 3. Developing efficient diagnostic and treatment-response predictive markers that help avoid unnecessary administration of therapy to patients that are less likely to benefit from a specific option of therapy. However, these goals are far from being achieved given the complexity of the disease, given the large number of genetic and epigenetic factors that have potential influence on the phenotype and its response to therapy. Today, estimated 25000 known protein-coding genes, non-coding RNAs, more than 250,000 proteins and epigenetic factors – are the basic variables. In addition, interactions between these proteins, feedback mechanisms, mutations, and copy number variations, combinatorial transcription factor binding, and epigenetic modifications– are other variables. How to identify those molecular alterations in the high-dimensional omics-universe that can significant influence on the phenotypic consequences? How to effectively predict treatment response? Microarray technology has been the most evolved, widely used and robust technology. MicroArray Quality Control (MAQC) project has helped to improve the interplatform consistency of microarray data and

has facilitated the merger of publicly available datasets from different technology platforms and different cohorts, for performing studies to infer novel classes and thereby to generate the clues of class-specific diagnostic and therapeutic markers. High-throughput methods including microarrays generate multivariate and multidimensional data. Several dimensionality reduction methods have been developed in the past decade to infer the reduced geneset that represent the maximum variance within the data and can be used for any further analysis such as biomarker search, class-description or class-inference. Given the tens of thousands of covariates (expression measures of genes) and small number of samples in a typical expression dataset to capture the heterogeneity of genes, overfitting and false positivity are likely limitations regardless of the strength of statistical methods applied. One of the possible solutions is to reduce the multidimensional structure of data by not analyzing the individual genes, but sets of genes representing the biological processes. Such an analysis would shift the focus from individual genes to the processes that they are involved in. Thereby the effect of the variability and noise of individual genes could be reduced and the sensitivity of analysis would improve because individual gene measures are weighted by the overall changes of genes within the pathway or process. In a realm of individualized diagnostics and therapeutics, activity status of biological pathways and networks is the key information for planning the targeted therapy. Therefore pathway-based analysis has immense potential to bridge the gap between the genomics and cancer management.

The work presented here as a part of this thesis starts with a study on potential transcription factors linked to the molecular subtypes, advances to study pathway deregulation by p53 mutation status and demonstrates the prognostic impact of p53 signaling and VEGF expression status. The key finding of the work is that over-expression of VEGF mRNA is an important predictor of survival in breast cancer, remarkably in a group of patients categorized otherwise as having favorable prognosis. The work then investigated VEGF expression class-specific miRNA-mRNA modules by using the miRNA and mRNA expression profiles from the same patients.

Index

Papers included in the thesis

Paper I:

Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes.

Joshi H, Nord S, Frigessi A, Børresen-Dale AL, Kristensen V.

BMC Genomics. 2012 May 22;13:199. *doi: 10.1186/1471-2164-13-199*.

Paper II:

Potential tumorigenic programs associated with *TP53* mutation status reveal role of VEGF pathway.

Joshi H, Bhanot G, Børresen-Dale AL, Kristensen V.

Br J Cancer. 2012 Nov 6;107(10):1722-8. doi: 10.1038/bjc.2012.461.

Paper III:

Implications of *VEGFA* upregulation on microRNA-mRNA Modules in Breast Cancers.

Joshi H, Børresen-Dale AL, Kristensen V.

Manuscript

List of abbreviations

Akt	Serine/threonine-specific protein kinase
CK-5/6	Cytokeratin 5/6
CMF	Cyclophosphamide, methotrexate, and 5-fluorouracil
CYP450	Cytochrome P450
DCIS	Ductal carcinoma *in situ*
DMFS	Distant metastasis-free survival
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
FISH	Fluorescence *in situ* hybridization
GCDFP15	Gross cystic disease fluid protein-15
Her-2/neu	Human Epidermal Growth Factor Receptor 2/Neu
ID (/L) C	Infiltrating ductal (/lobular) carcinoma
IGF1R	Insulin-like growth factor 1 receptor
IHC	Immunohistochemistry
LCIS	Lobular carcinoma *in situ*
MDR1	Multidrug resistance protein 1, ATP-binding cassette, sub-family B (MDR/TAP), member 1
MINDACT	Microarray In Node-negative Disease may Avoid Chemotherapy
PARP	Poly ADP ribose polymerase
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
PR	Progesterone receptor
qRT-PCR	Quantitative real time polymerase chain reaction
RTK	Receptor tyrosine kinase
RT-PCR	Reverse transcription polymerase chain reaction
SAFE	Significance Analysis of Function and Expression
TAILORx	Trial Assigning Individualized Options for Treatment (Rx)
TF	Transcription factor
TFAC	Paclitaxel (Taxol), 5-fluouracil, doxorubicin (Adriamycin) cyclophosphamide
VEGF and VEGFA	Vascular endothelial growth factor
USFDA	United States Food and Drug Agency

List of Figures and Tables

1 Overview

Breast cancer is the leading cause of cancer and cancer mortality among all cancers in females with an estimated incidence of 1.38 million globally and 458500 deaths (equivalent to one in every seven cancer deaths) worldwide among women in 2008 [1]. With the annual percentage change of 0.1% in incidence between 2004-2010, the recent estimates in the United States show that one in every eight women (estimated 12.38% women) carries the risk of being diagnosed breast cancer during their lifetime [2]. Over 1.1 million cases of breast cancer are diagnosed across the world each year, compared with about 500,000 cases in 1975. This represents about 10% of all new cancer cases and 23% of all female cancers. An annual prevalence of more than 4.4 million cases of breast cancer is expected worldwide by the year 2012, with its occurrence in 3 out of every 10 females during their lifetime worldwide and likelihood of one of them to die. Incidence varies considerably across the world ranging from 19.3 per 100,000 in Eastern Africa to 89.7 per 100,000 in Western Europe [1]. The differences in the access to treatment world-wide mainly contributes to the fact that despite of the wide-ranging differences in incidences, breast cancer remains the most frequent cause of cancer deaths (12.7% of total) among women of both developed and developing world. Recent estimates in the United States show the mortality decreasing at the rate of slightly below 2% annually. Decreasing mortality can be attributed to early detection, screening programs, use of predictive and prognostic markers, introduction of Her2-neu targeted therapy, use of adjuvant (particularly post-operative Tamoxifen). However, the observed reduction in mortality is suboptimal compared to the size of likely to be benefited patient groups identified by available predictive markers. Besides affordability and access to treatment options such as Trastuzumab, fewer drugs compared to the broad spectrum of biological complexity and heterogeneity – remains to be very important factors underlying the suboptimal improvement in mortality figures and in relapse-free patient survival.

An important priority to effectively reduce the mortality from breast cancer is to advance the expertise in molecular diagnostics and therapeutics and to translate it into routine breast cancer management practices with specific emphasis on those molecular classes of cancer that are likely to recur and experience poor survival. Not just the advancement of such expertise and its translation into practice but priority is also required in making these options uniformly accessible, affordable and cost-effective across the world.

1.1 Complexities underlying the molecular portraits of breast cancer

The concept of molecular subtypes of breast cancer has constantly evolved since the last decade in an effort to understand the heterogeneity and resultant phenotypic diversity of breast cancer. Molecular heterogeneity can be within-tumor heterogeneity (such as cellular heterogeneity, heterogeneity of molecular programs active within different cells within the same tumor, etc.) in the tumor microenvironment and between-tumors (from different patients) heterogeneity. Together these two types of heterogeneity form major challenge to the molecular categorization and its successful application to personalized medicine. The existing description of molecular portraits can be viewed as a categorization derived by the efforts to understand between-tumor heterogeneity. Originally defined molecular portraits in the past decade, that are based on differential gene expression pattern demonstrated in the microarray data, have been further studied by the aberrations at the level of DNA methylation, microRNA (miRNA) because of the technological advances. In parallel with the increasing understanding about the molecular portraits, advances in understanding the cancer stem cells have led to the discussion about their role in initiation, maintenance, progression and recurrence. While comparing to the conventional categorization of breast cancers, which is mainly based on grade, stage, size, histopathology, categorization of breast cancer based on molecular heterogeneity could provide more detailed explanation of the phenotypic diversity because of being a better reflection of the biological differences. Thorough understanding of molecular heterogeneity and corresponding molecular categorization may implicate a paradigm shift from conventional diagnostic and therapeutic protocols and more precise prognostic profiling. However, the success of any type of molecular categorization including the molecular portraits depends upon how closely these molecular categorizations represent the differences of inherent biological complexity observed in breast cancer.

1.1.1 Motivation underlying the study of molecular profiles of tumor

The following factors form the major motivating factors for the study of molecular profiling in breast cancer.

1. Tumors with relatively similar grade, stage, size may have different biological profiles.

2. Tumors classified into same morphological class on the basis of histological study may vary considerably at molecular level and therefore morphological classes are insufficient to represent the biological differences.

3. Different molecular profiles indicate differences in diagnostic and therapeutic markers. While targeted interventions might achieve relatively higher specificity of action compared to the conventional therapeutics, characterization of robust and standard molecular classification could significantly improve the patient outcome.

1.1.2 Reasons for the limited clinical utility of molecular portraits

As mentioned earlier, the successful application of molecular classification depends on how well it represents the molecular heterogeneity and the distinctive biological character of the tumor. With this viewpoint, the following are the limitations of existing molecular portraits:

1. Lack of uniform and standard definition of known molecular portraits

2. Lack of practical and cost-effective diagnostic methods that can be used at clinical level to classify each cancer into molecular portraits

3. The existing molecular portraits are broadly defined classes with considerable within-class heterogeneity. For each molecular portrait there could be uncharacterized biological differences. This core limitation of existing molecular portraits questions the marginal benefit of applying molecular portraits beyond the conventional protocol used for diagnosis and for treatment response prediction.

4. Molecular portraits are the snapshots mainly based on statistical approaches, such as clustering, classification and differential expression. Variation presented by these methods in high-throughput data represents statistical measure of gene-expression differences. The actual differences in their biological effect-sizes may not necessarily be proportionate to the expression differences as estimated by these methods.

5. Results may vary by the high-throughput technological platform used and by laboratory where samples were processed and therefore the interpretation of molecular classifications may vary. This is one of the general limitations in understanding the molecular heterogeneity.

1.2 The histopathological classification

The conventionally used scheme of classification is based on the grade and the differences in architectural features and growth patterns identified by histopathological study of the tumors. The grade of tumor is based on the degree of pleomorphism, loss of tubule formation, etc. and thus provides an idea about the aggressiveness of the tumor. Histological typing is based on the cytological and morphological patterns of tumor. Several studies have shown that histological grade can be used as an indicator of survival. Both grade and histological type provide complimentary information.

1.2.1 Histopathological classes of breast cancer

Invasive ductal cancer - *not otherwise specified* (NOS) is the most common histopathological subtype with its occurrence in 40-75% of invasive breast cancers. Because of having no peculiar histopathological feature, it is often referred as- *not otherwise specified (NOS)*. Degree of differentiation can vary from well-differentiated with abundant gland formation to poorly differentiated having sheets of cancer cells. These tumors are less common in younger age group. Prognosis is intermediate.

Invasive lobular cancer is another histological subtype with occurrence of about 10-15%, typically presenting with grade 2, lack of cellular cohesion and more likely to be multicentric and 20% chance of bilateral presentation. Nuclei have typically signet ring appearance because of round or oval shaped notched nuclei with thin rim of cytoplasm. About 10% of the lobular tumors present with grade 3 pleomorphic features and clinically aggressive behavior.

Medullary carcinoma is a relatively less frequent subtype with 1-5% of invasive cancers, usually presenting as a syncytial growth, marked with stromal infiltration by lymphocytes and plasmocytes. Regardless of the high proliferation and poor differentiation, this subtype often carries a good prognosis.

Tubular carcinoma is another infrequent subtype with about 2-5% of invasive breast cancers. The histopathological characteristics are marked by high degree of differentiation, randomly distributed cells in tubular architecture and open lumens, small size of cells, scanty mitosis and low degree of pleomorphism. These tumors have favorable prognosis.

Cribriform Carcinoma is a rare subtype (1 – 3 %) with favorable prognosis. The histopathological features are cribriform architecture, scanty mitosis, and low to medium degree of pleomorphism.

Mucinous or colloid carcinoma is a subtype marked by the uniform small cells with eosinophilic cytoplasm surrounded by extracellular mucus. Other characteristics are – lack of myoepithelial cells, low degree of pleomorphism and scanty mitosis. The tumor is seen in patients above the age of 60.

Apocrine carcinoma is a subtype that arises from the apocrine cells of sweat glands of breast. Apocrine cells presenting with: abundant cytoplasm, vesicular nuclei, GCDFP15 positivity and apocrine snouts appearance (secreted granules in the apical cytoplasm).

Micropapillary carcinoma is an aggressive but uncommon subtype with poorly differentiated cells with prominent nucleoli, coarse chromatin, and increased mitotic count and higher likelihood of lymph node metastasis.

Other rare varieties include metaplastic carcinoma, lipid-rich carcinoma, glycogen-rich carcinoma, adenoid cystic cancer and inflammatory carcinoma.

1.2.2 Pros and cons of histological classification

The following are the advantages of histological classification of breast cancer:

1. Histopathological subtypes are practically feasible and have proven to be a cost-effective, gold-standard and routinely accepted method for diagnosis of invasive cancer.

2. Some of the specific subtypes are also able to predict the prognostic profile based on histopathology. This can be used to supplement the information obtained by IHC or molecular studies.

3. Histological grading system can provide a criterion for deciding the need for post-operative chemotherapy.

4. Histopathological response can be a possible mean to evaluate or monitor the response to treatment. While it is more suitable in case of clinical trials, repeating the biopsy has no proven value in routine cancer management.

The following are the limitations of histological classification:

1. Most histological subtypes cannot specifically indicate any particular biological feature. That means the histological subtypes do not have predictive utility.

2. While it is possible to gain certain prognostic indications based on the histopathological study, such prognostic stratification is broader compared to the one provided by molecular classification.

3. Histopathological appearance might sometimes lead to differing conclusions. Even expert pathologists might have differences of opinions.

4. Unsuitable for monitoring the treatment response, given the invasive procedure

5. Large fraction is categorized as grade 2. Besides most tumors have the histological type of IDC-NOS. Therefore the information gain from such classification is limited.

1.3 IHC in breast cancer

IHC has also got a vital role in diagnostics, prognostics and predicting the response to therapy. Conventionally, the histological classification together with IHC-based markers has been used in determining the management strategy of breast cancer. ER has been the oldest known prognostic and predictive marker, even before the IHC came into practice in 1990s. While the decrease in breast cancer mortality is observed over past few years, use of adjuvant therapy, particularly post-operative Tamoxifen adjuvant therapy- is an important underlying factor. ER status (together with PR status) helps in predicting which patients would likely benefit from Tamoxifen. Earlier ligand-binding assays for assessment of ER status have been replaced by IHC. Today most experts worldwide recommend both ER and PR measurement in all primary invasive breast cancers (but not in DCIS) to identify the patient subset likely to benefit from the hormonal treatment in both the adjuvant and metastatic settings. PR is a co-dependent marker with ER. Positivity of both ER and PR has been shown to improve the accuracy of likelihood of responsiveness to endocrine therapy. In addition to the tremendous

evidence regarding the clinical value of ER and PR assay, Her-2/neu positivity has also been shown to confer poor prognosis in breast cancer. With the introduction of Herceptin™ (trastuzumab) since 1998, Her-2/neu became the predictive marker for responsiveness to Herceptin therapy. The overexpressed Her-2/neu antigen in tumor is targeted by humanized monoclonal antibody (Herceptin). About 20-30% of patients show Her-2/neu overexpression. IHC is one of the standard tests available for Her-2/neu protein assay. FISH is an alternative to IHC, though it is indicated for improving accuracy when tumors score 2+ by IHC. IHC provides a score representing the degree of positivity of Her2 protein, whereas FISH provides the status of Her2 gene amplification in nucleus. It has to be noted that the predictive value of ER, PR and Her-2/neu is not merely reflected by the positivity or negativity on IHC staining, but also by the quantity of antigen present. IHC-based scores of these markers that represent the degree of positivity or negativity can be useful in combination with histological grade and type –for improved clinical decision-making.

1.4 Combining the histopathology, IHC and molecular portraits in the clinics

Conventionally diagnostics is based on histopathology. In order to determine the indication of chemotherapy, clinicians conventionally rely upon criteria, such as size, grade, Ki-67 index, ER/PR/Her-2 status and nodal involvement. These criteria largely provide an idea about the aggressiveness, proliferation, hormone receptor status etc. However, these criteria are insufficient for determining the indication of chemotherapy and efficient response prediction for the available options in view of the heterogeneity and variation in response.

Given the considerably high proportion of non-/partial responders to chemo-/endocrine/targeted therapy, the vital question is to what extent the inclusion of available molecular knowledge in breast cancer management can improve the patient outcome, proportion of responders to chemotherapy and help avoiding the unnecessary chemotherapy to the potential non-responders. Molecular portraits and the corresponding commercial assays can improve the understanding of the biology of tumor in individual patient and can provide an opportunity for more informed choice to clinicians in optimizing the plan of treatment of individual patients.

It is known that ER+ve group has a poor responsiveness to chemotherapy compared to triple negative groups. Relative advantage of using molecular portraits in predicting treatment response was demonstrated by similar higher rates of complete pathological response achieved with neoadjuvant chemotherapy even after exclusion of triple negative patient group by using the 70-gene signature [3]. Besides the assay based on 70-gene signature separates the patients with nodal involvement (up to 3 nodes) and excellent prognosis from the rest [4].

Question is – can we use the existing definitions of molecular classes? If so, how to utilize this available knowledge of molecular portraits for improving the clinical decision-making? There has been availability of commercial assays that can be used in combination with the routine practice of histopathological assessment and IHC. *Table 1* shows a number of assays that can be useful together with histopathology and IHC.

The prediction of prognosis and response to chemotherapy, inclusion of molecular classifications might also help response prediction to the targeted therapy. The need of precise diagnostic and predictive tools is evident as therapeutic advances aim to target a specific biological marker or a pathway. One such example is Her-2 overexpression and response to trastuzumab. Only about 30% of the tumors respond to Trastuzumab therapy among potential target group of patients with Her-2 amplification detected by IHC or FISH as a criterion for therapy. There is a lack of precise assays for the response-prediction to Trastuzumab. In case of molecular classes, such as basal-like, normal-like have no unique markers of biology that can indicate a response to any specific drug.

This means that the existing definitions of molecular classes need to advance and more precise tools and assays have to be developed to improve the predictive, therapeutic and prognostic performance.

1.4.1 Commercially available assays for diagnostics

Some of the commercially available assays are listed in *Table 1*. MammaPrint is the first assay that is approved by USFDA's new *In Vitro* Diagnostic Multivariate Index Assay (IVDMIA) classification. Many of these assays based on characterizing molecular profile of breast cancer are not shown of significant clinical value based on large-scale public trials. Besides, higher cost of implementing them at clinical level has raised the concern among health economists. But the most important strength of these assays is the marginal utility and

improvement in clinical decision-making relative to IHC and histological review of tumors. For the time, two large-scale trials have been implemented– MINDACT [5] and TAILORx [6] for evaluating MammaPrint and Oncotype Dx, respectively. The MINDACT trial has been recruiting about 6000 patients with invasive, node –ve, stage 1, 2 or 3 breast cancers. In this prospective cohort, the trial aims to compare two groups of patients – group I: low genomic risk and high clinical risk; group II: high genomic risk and low clinical risk. In these groups, cases with high clinical or genomic risk respectively - would receive the chemotherapy and the study would confirm that cases with low genomic risk and high clinical risk could be safely spared chemotherapy without influencing the DMFS. Besides, this trial would also help in inference and validation of novel multigene signatures that can predict response to chemotherapy and endocrine therapy. The TAILORx is organized by the National Cancer Institute to test in a prospective cohort to evaluate the utility of Oncotype Dx in determining whether diagnosed ER+ve breast cancer cases with intermediate recurrence score of Oncotype Dx would benefit from adjuvant chemotherapy or not. This trial has been recruiting ER/PR +ve, Her-2/neu –ve, lymph node –ve breast cancer cases.

1.4.2 Correspondence between IHC and molecular portraits

There have been a number of studies that tries to simplify the criteria of defining the molecular portraits on the basis of IHC. It is not certain to what extent the surrogate IHC-based markers can reflect the underlying biological traits represented by the molecular portrait. Studies that have discussed the correspondence between the IHC and molecular portraits are shown in *Table 2*, including the latest study[7] that provides comprehensive discussion in it.

While gene expression profiling based molecular portraits are not practically suitable for routine use in clinics, IHC has far more proven to be practical and cost-effective means as a method for use in clinics. So far limited set of IHC-based markers are used to describe the molecular portraits or to predict the response to chemotherapy and prognosis. While the knowledge of molecular portraits is evolving, it is crucial that IHC continues to evolve in terms of its applied value by inclusion of more IHC-based markers.

Table 1: Commercially available multigene signature-based tests for predictive and prognostic purposes

	Gene-set size	Method/ Technology	Indication
MammaPrint [3, 8]	70	Oligonucleotide array	Prognostic: predicts the recurrence risk in both ER+ and ER – cases
Oncotype Dx [9]	21	Quantitative RT-PCR	Predictive for response to tamoxifen and to the CMF adjuvant chemotherapy regimen for ER+ cases, either stage I or II node –ve or postmenopausal node +ve; can also be prognostic
The Rotterdam Signature [10]	76	Oligonucleotide array	Prognostic: predicts the risk of distant metastasis in node –ve cases irrespective of the ER status
The Invasiveness Signature [11]	186	Oligonucleotide array	Prognostic: Predicts the risk of metastasis and poor survival in all breast cancers irrespective of the ER/node status
AmpliChip CYP450 Test [12]	2	Oligonucleotide array	Predictive: Determines the genotype of CYP- 2D6 and 2C19. Used in ER+ve cases to evaluate Tamoxifen sensitivity
NuvoSelect [13, 14]	30 and 200	cDNA array	Predictive: predicts response to preoperative TFAC chemotherapy; Prognostic/predictive: predicts outcome after 5 years of endocrine therapy
Wound response signature [15]	380	Oligonucleotide array	Prognostic: For risk stratification
Celera Metastatic score [16]	14	RT-PCR	Prognostic: predicts recurrence risk in ER+ve, node–ve caes treated with Tamoxifen
Breast bioclassifier [17]	50	qRT-PCR	Prognostic: Predicts risk in both ER+ve and ER–ve cases
Breast Cancer Two-Gene Expression Ratio (H/I™) [18]	2	qRT-PCR	Prognostic: predicts 5-year recurrence risk in ER+ve, node negative cases
eXagenBC [19]	6	FISH	Prognostic: provides a prognostic index

Table 2: Table shows IHC-based status of known markers and corresponding molecular portraits, as defined by various literature sources

IHC-based status of markers	Corresponding Molecular portrait	Source
ER–, Her-2– or low, CK-5/6+ and/or EGFR+	Basal-like	Nielsen et al, 2004 [20]
ER−, PR−, Her-2−, CK-5/6+	Basal-like	
Her-2+, ER−, PR−	Her-2+	Carey et al. 2006 [21]
ER+ and/or PR+, Her-2−	Luminal A	Spitale et al.2009 [22]
ER+ and/or PR+, Her-2+	Luminal B	
ER+, PR +, Her-2 –, and Ki67 index<14%	Luminal A	
ER+, PR +, Her-2 –, and Ki67 index≥14%	Luminal B	Cheang et al. 2009 [23]
ER+, PR +, Her-2 +	Luminal Her-2+	
ER+, PR+, Her-2−, CK-5/6 or EGFR-	Luminal 1 (Luminal A)	
ER+, PR+, Her-2−, CK-5/6 or EGFR+	Luminal 1 (Luminal B)	
ER+, PR+, Her-2+, CK-5/6 or EGFR+ or –	Luminal 2 (Luminal B)	Blows et al. 2010 [7]
ER–, PR–, Her-2+, CK-5/6 or EGFR+ or –	Non-Luminal Her-2+	
ER–, PR–, Her-2–, CK-5/6 or EGFR+	Triple Neg : Core basal	
ER–, PR–, Her-2–, CK-5/6 or EGFR-	Triple Neg : 5-Negative	

1.4.3 Correspondence between the histopathology, IHC and molecular portraits

It could be intriguing to compare the molecular heterogeneity to IHC and histological classes provided by grade and type. Because differences in molecular events that underlie the causation and progression of cancer could give rise to differing morphological patterns. Besides the molecular heterogeneity can also determine the degree of differentiation of particular clones of cells. Specific driver mutations or genetic abnormalities have been known to confer selective growth advantage under a specific set of selective pressures, thus evolving into specific clonal dominance and proliferation, reflected in the tumor grade. As a result of several mechanisms – observations made at histopathological level, levels of biomarkers as determined by the IHC and the snapshot of differential gene expression patterns representing the molecular portraits – are all linked and show corresponding differences to certain extent.

Figure 1 shows how these different types of classifications have correspondence in between one another. Most high-grade tumors are classified as basal or ERBB2+/Her-2+ at molecular level. The status of biomarkers by IHC of these tumors is as shown in *Table 2*. Histological appearance of these tumors shows mostly cells with higher tumor grade, lower differentiation. Histological subtypes high-grade ductal, medullary, metaplastic cancers correspond to these tumors. Association of medullary subtype with triple negativity and BRCA1 germline mutations [24] and expression of cytokeratins and EGFR correspond well with the basal-like group [25]. Tumors described as luminal B at molecular levels are intermediate grade tumors and at IHC level they represent largely as ER+/PR+ but some of the tumors might be Her2+. Pleomorphic variety of lobular, Micropapillary and apocrine tumors correspond to this group. Regarding the patient outcome, micropapillary variant has been reported to have high proportion of tumors with ER and Her2 positivity [26] and though not associated with significant difference in patient outcome compared to ductal cancers with similar nodal status [27]. Molecular class luminal A is usually ER+ /PR+ and Her-2– and usually associated with good prognosis. Correspondingly, large majority of ductal carcinoma- NOS, tubular [28], mucinous [29] and classical lobular and cribriform [30] carcinoma share similar IHC and prognostic profile. About 70–95% of lobular carcinomas are ER+ [31] having low Ki67 index

[32] and with the exception of pleomorphic variety, Her2+ [33] and p53 mutations [34] are less frequent compared to the ductal cancers.

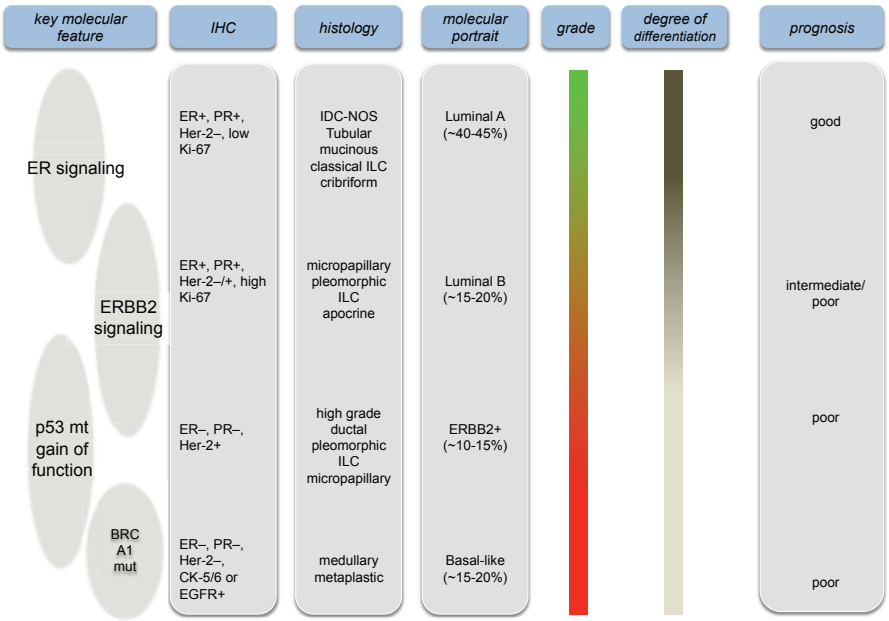


Figure 1: Possible overlap between molecular, histological and IHC-based class.
Comparison of the molecular portraits to the histological and IHC based classes together with the grade and degree of differentiation

1.5 Pathway approach in breast cancer

For any form of categorization of breast cancer to be sensible, the most important criteria is – how uniformly the core biological characteristics are represented in each of the defined classes. The objective is to link the mechanisms of carcinogenesis that involve processes and pathways causing and driving the cancer process to the diagnostics, therapeutics and prognostics, while minimizing the within-class heterogeneity. This means characterization of classes that have unique and class-specific differential activation or repression of specific biological pathways and processes responsible for driving the specific categories or classes of cancer. This is becoming more important priority with the expanding world of molecular therapy. Application of pathway approach in breast cancer implicates the methodological

means for quantification of the pathway activity in each individual tumor. Inference of pathway activity is performed by a variety of approaches. For instance, pathway activity can be shown as a probability [35] of activation or as a summative pathway activity score [36], based on consistency of the pathway-specific genes' differential expression. The key objective of the statistical approach is to predict the key pathways out of many, having dominant role in cancer-progression specific processes. Cancers showing such specific pattern of pathway perturbations should be categorized in one particular subgroup that can likely respond to the pathway-targeting therapy. The more specific the pathway identification is, the higher likelihood of such therapeutic options to prove efficacious while minimizing the chances of relapse and resistance. The implication of such effort can be predicting the likelihood of resistance and recurrence in a group of cancers that are broadly described to have good prognostic profile based on the conventional diagnostic protocols or vice versa. Classic example is – only half of the hormone receptor positive breast cancers respond to Tamoxifen [37]. Among the non-responders fraction of the cases have dominant activity of other cancer progression-related pathways (with the exception of those cases with ESR1 mutation).

1.5.1 Pathways concept in context of cancer genomics

Pathways are defined as a set of functional interactions between the genes, proteins or other molecular components that together act and thereby perform a specific biological process. Pathways can be categorized as: signaling pathways, metabolic pathways and disease-associated pathways. The disease-associated pathways are the set of interactions found to be functional in certain disease or disease subgroup. The concept of pathways makes it convenient to formulate network models of genes and proteins involved in specific pathways and then to perform systems modeling of a particular pathway or of a set of pathways. Besides it also helps in understanding the interaction between the pathways and models the possible consequences.

Even when pathway concept provides the simplified means to understand the phenotype, it is important to note that involvement of pathways in cancer is a dynamic process. Because cancer is a multistep process, where driver mutations initiate the cancer by altering the one or more pathways, and eventually more genes might accumulate mutations that can alter their function and can influence the function of the downstream genes. This means that

perturbations of pathways demonstrated at a particular time-point represent only a snapshot of pathway activity, not as an ongoing process.

1.5.2 Publicly available pathway databases

Pathway databases are the repositories of the available experimental or sometimes prediction-based evidence of gene-gene, gene-protein, protein-protein or other forms of interactions organized by commonality of the processes they are involved in. The utility of pathway databases is not merely limited to the curation of the available interaction data in pathway format but these databases provide a standard protocols of data-exchange between other relevant databases or tools, serve as a means for statistical and graphical approaches of pathway analysis and network modeling. In a pathway-based approach, functional groupings of genes that are based on canonical pathways curated from literature resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [38], Reactome [39], National Cancer Institute's Pathway Interaction Database (PID) [40], curated functional genesets(C2) of MSigDB [41], Gene Ontology [42], PantherDB [43] or other experiment-based annotations describing the interactions between the genes. Reactome and C2 of MSigDB cross-reference with several other databases and thereby provide better inclusion of available evidence. Many of these databases provide support to multiple data formats, such as Biological Pathways Exchange (BioPAX) [44], Systems Biology Markup Language (SBML) [45], KEGG Markup Language (KGML) [38] .

1.5.3 Advantages of pathways-based analysis over individual gene-based analysis

Cancer genome is usually characterized by derangement of several biological processes as a consequence of altered function of genes and proteins. The following are the advantages of pathway-based analysis over the individual gene based analysis.

i. Genes act in concert to activate or repress specific pathways. Genes can be involved in one or many pathways. Subtle changes in the expression of one or more genes and their complex interactions can strongly alter the activity of the process or pathways and thereby can shape the biology underlying a specific disease or cancer subtype [46].

ii. Biological pathways are altered as a consequence of a variety of defects of individual genes involved in the pathway or their regulators. This means similar phenotypic manifestation of cancer can be a result of one of the many possible genetic or epigenetic alterations. Pathway-supervised approaches can help understand the basis of such alterations by incorporating interactions of genes involved in same or related pathways. The best example is the p53 pathway, which is inactivated mostly via p53 inactivating point mutation. Notably inactivation of p53 signaling pathway can also occur by alternative mechanisms such as *MDM2* amplification or MDM2 splice.

iii. Certain gene mutations are more frequent compared to others and can alter a set of protein-protein interactions, giving rise to alteration of a process or pathway. Others are infrequent mutations or epigenetic modifications giving rise to rare forms of cancer. Application of pathway approach groups the cancers by pathway and not by individual gene alterations. Thereby it increases statistical power for analyzing the biology of uncommon genetic alterations.

iv. From methodological perspective, methods used to identify differential expression suffer from the major setback of being dependent on the most suitable statistical cut-off that can identify most functionally altered genes. Statistical significance of differential expression values might not necessarily represent the biological significance. Besides, methods used for quantifying absolute gene expression levels such as microarrays, RNAseq have their own limitations. Therefore changes in gene expression values that might not pass the cut-off of statistical significance, will remain undetected (false negatives). Pathway-based approach can be used as an improvement for biomarker search [47, 48].

1.6 Deregulation of transcriptional networks in cancer pathways

Gene transcription is a process determined by the complex interaction of one or more regulatory transcription factors with the putative regulatory region on a gene promoter. It is also known that genes that are co-expressed are likely “co-regulated”. This means the group of genes involved in a given biological process might be regulated by a set of common transcription factors (TFs) and therefore can share a set of corresponding transcription factor

binding sites (TFBSs) for allowing the binding of their regulator TFs. The combinatorial effect of multiple transcription factors binding the promoter of the given set of genes could be induction or repression of target gene transcription. The regulatory binding by transcription factors is a context-specific event and is selective to the specific target promoters. This regulatory mechanism maintains the homeostasis in the signaling pathways activity and thereby regulates the cell physiology.

Deregulation of transcriptional networks within the biological pathways can occur as a consequence of the alterations of upstream regulatory TFs, alterations in the co-activators of signaling cascade, elimination of negative regulatory feedbacks or by alterations in the downstream signal transduction pathway. The alterations of transcription factors occur due to mutations, deletions, amplifications, or due to post-transcriptional modification. Certain alterations of TFs might confer oncogenic properties to cells by perturbing the downstream processes involved in proliferation and growth regulation, DNA repair and replication. Alterations of TFs can be linked to the specific sets of target genes and pathways that are likely to be perturbed in subclasses of cancer. This implies that inference of molecular phenotype-specific regulatory TFs is of immense importance in developing gene-based diagnostic and therapeutic strategies.

1.6.1 Inference of transcriptional factors underlying deregulation

Inference of regulatory transcription factors can be performed experimentally (i.e. ChIP-sequencing, ChIP-chip) or by using *in silico* methods.

Here the basic rationale of the *in silico* approach has been briefly described. The methodological approach for the computational inference of potential regulatory transcription factors underlying molecular subclasses is the following:

Molecular class-representative clusters in the gene expression signatures are often composed of a set of co-expressed genes observed only in a subset of cancers. One strategy to find the potential functional transcription factors in a given cancer class is to find a set of significantly over-represented motifs in the promoters of a set of co-expressed genes (usually a signature genes of a given phenotype class). Computational methods either search for a known TFBS motif or for new, previously uncharacterized motifs (*de novo* motif discovery). The *de novo* motifs can be filtered based on the criteria such as degree of conservation of the motif across

species. Discovered potential TFBS motifs needs to be experimentally validated. Among the gene promoters that show statistically significant overrepresentation of TFBS, those that show significant co-expression of their corresponding TF gene-target gene pairs in the expression profiles of given cancer class- increase the likelihood of true positivity of biologically functional interaction.

Variations in transcriptional deregulation form an important source of heterogeneity within molecular classification. For example, hormone receptor negative breast cancers might be composed of the samples having mutations in p53, PIK3CA, BRCA1 etc.

1.6.2 Defining novel cancer classes by the activity of transcriptional hubs

Class-representative metagene consists of genes including the ones that code transcription factors. When a particular transcription factor regulates multiple target genes involved in diverse processes and pathways that are involved in more than one class-representative clusters, it is referred as transcriptional hub protein (T_H). Given the multitarget interactions, it appears as a topological hub in the disease-specific functional interaction networks. Dysfunction due to under-/over-expression, amplification, deletion, mutation of these T_H genes and resultant aberrant activity of TF protein might have diverse consequences on the expression of genes regulated by it and thereby can influence activity of all connected pathways. Co-existing mutations or aberrant expression of other genes might act as an additional source of heterogeneity in cancer. These hub genes might also have conserved function across species and are often linked to chromatin modifications [49].

In *figure 2*, $T_{H,}$ the hub transcription factor is shown to have multiple targets in multiple class-defining clusters (s1-s4). Given the larger impact of the differential activity status of hubs, corresponding novel cancer phenotypic classes can be defined. These classes might not just represent the differential activity of TF hub and its target genes, but can have considerable clinical relevance.

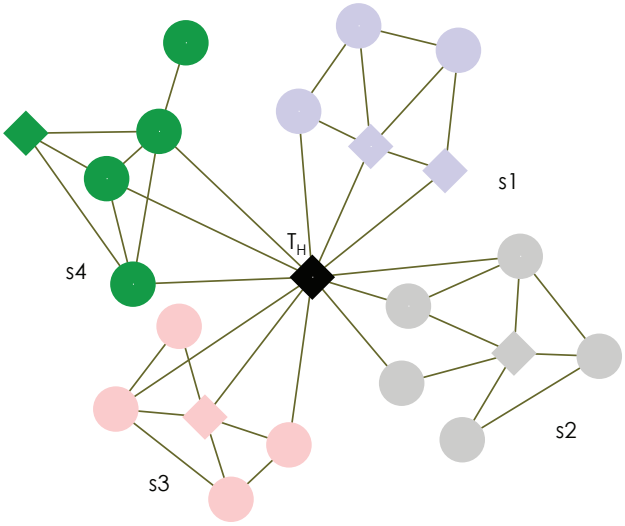


Figure 2: Graphical illustration of the transcriptional and topological hub protein
Transcriptional hub (central node in black) regulates the target genes (circular nodes) including transcription factor genes (diamond shaped nodes) in four class-defining clusters (s1-4 in different colors).

1.7 Transcriptional deregulation by p53 in breast cancer

TP53 is a tumor suppressor transcription factor with paramount clinical value because of its association with tumor progression [50], metastatic potential [51], early relapse [52], response to chemotherapy [52, 53], ultimately to prognosis and survival [54, 55] . TP53 is a master regulator transcription factor, which is involved in key processes such as cell cycle, DNA repair and genomic stability and most importantly also of cell death. The core of p53 functions is by the sequence-specific DNA binding to target genes that are involved in carrying out diverse cellular functions. Recent studies have revealed its role in differentiation [56], angiogenesis [57], mitochondrial respiration[58], glycolysis [59], glutamine metabolism [60], response to anti-oxidants [61]. While the levels of p53 are maintained low by the binding of MDM2, COP1, PIRH2 or JNK etc-mediated degradation in ubiquitin-proteasome proteolytic pathway, the activation of p53 occurs in response to the DNA damage and other types of stresses. The activity of p53 is initiated with the elimination of negative feedback

mechanisms including MDM2, increased mRNA translation of p53 and by increased transcription induced by specific modulators.

1.7.1 Determinants of functional specificity and promoter selectivity of p53

The following factors determine the functional selectivity of p53:

1. Its differential affinity to the response elements located on the target gene promoter

2. Differences in the post-translational modifications within amino-, carboxy-terminals or DNA binding domain

3. Alterations by the cofactors

In breast cancer, the p53 transcriptional program may get deranged because of one or more factors. Large majority of breast cancers are ER+ve and tend to have wild-type p53, whereas about 20-30% of them are associated with mutant p53. While deregulations in the transcriptional program are obvious in the presence of mutant p53 protein, deregulated transcription programs of p53 target genes can also occur with wild-type p53 in breast cancer. Recently transcriptional regulation by ERα and its consequences on transcriptional response of p53 target genes have been studied. ER can bind to p53 targets and thus can inhibit the p53-mediated apoptosis [62, 63]. This effect is also explained by the binding of ER that represses p53 on the p21, survivin, and MDR1 promoters with subsequent inhibition of gene expression. Besides, ERα can directly bind p53 and thereby can access its target gene promoters and may repress p53 transactivation by recruiting NCOR, SMRT, and HDAC1[64]. The ER effect on p53 transcription program can be partly explained by the observation that response to Tamoxifen-therapy in ER+ve breast cancers is better in cancers with p53 wild-type compared to ones with mutant p53 [55].

Mutations in p53 are often of missense (point mutation caused by the replacement of a single nucleotide) variety, and occur frequently within the DNA-binding domain of p53 protein. Thus the deregulation of p53 transcription program by mutant p53 varies widely according to the location and type of mutation. For instance, codons 273 and 248- the mutation hotspots presenting with sequence alterations in the DNA-binding (contact) region of p53 can alter the sequence-based affinity of p53 to its targets, whereas mutations on codons 175 and 220 can

lead to structural alterations in the DNA binding region. Besides, most missense mutations can lead to partial or subtle effects on p53 transcriptional program and therefore the overall outcome on pathway activity and phenotype may vary considerably [65]. Many mutant p53 forms can induce cell cycle arrest but lose the ability to induce apoptosis [66]. However, some studies also propose that mitochondrial and cytoplasmic fractions of p53 may retain the apoptotic function regardless of mutation status of p53 and without the influence of domain negativity [67, 68]. Mutant p53 can also alter the binding of cofactors on the target genes. Some of the effects include induction of IGF1R [69]- that in turn can activate PI3K/AKT and MAPK signaling pathways [70], induction of VEGFA [71]-responsible for increased angiogenesis and invasion, induction of NF-κB activity in response to TNF-α [72]. Mutant p53 also gains new roles as transcriptional activator or repressor (gain of function). *EGFR, HSP70, MDR-1, VEGFR*- are some of the genes that can be transcriptionally activated by mutant p53 [73]. This results in chemo-resistance and activation of tumor promoting pathways- such as angiogenesis, proliferation and transformation.

Mutation status of p53 is not only prognostic, but also its effect on patient survival varies according to the ER status in breast cancer. Pathway analysis of breast cancer expression profiles aimed at investigating which pathways are the most significantly differentially enriched – identified at least 40 differentially enriched pathways by p53 mutation status. These pathways include metabolic pathways - such as glycine, serine and threonine metabolism, arginine and proline metabolism, sphingolipid metabolism; signaling pathways - such as p53 signaling, hedgehog signaling, calcium signaling, insulin signaling, MAPK signaling, ERBB signaling; and cancer pathways – such as renal cell cancer, pancreatic cancer, melanoma etc. Genes involved in ER signaling, PIK3K cascade, mammary gland development and apoptosis were found associated with wild-type p53 breast cancers, whereas genes involved in cell cycle, DNA replication, p53 signaling, purine nucleotide metabolism, p53 signaling and VEGF signaling were found upregulated or associated with mutant p53 breast cancer profiles [74].

1.8 Pathway-based molecular diagnostics

In a view of the dynamic nature of the biological processes and pathways that initiate and propagate cancer, the task of characterizing the activity status of molecular pathways in cancer is complex.

1.8.1 Characteristics of malignancy and pathway aberrations

Cancer cells have the property of hyperproliferation, invasiveness and metastasis to remote sites. Histologically, the following are the characteristics of cancer cell:

1. Increased nuclear/cytoplasmic ratio
2. Nuclear pleomorphism
3. Hyperchromatism and enlarged nucleoli
4. Bizarre appearance of mitotic spindle
5. Anaplasia or lack of differentiation

These histological features and their variations represent the manifestations of molecular and pathway aberrations. Characterization of the pathway aberrations that underlie these histological characteristics - is the mission of the pathway-based molecular diagnostics. The basic molecular traits or capabilities of cancer are earlier described as hallmarks of cancer [75] (shown in the *figure 3*). These ten basic traits are: evasion of growth suppressors, avoiding immune destruction, enabling the replicative immortality, tumor promoting inflammation, activating invasion and metastasis, induction of angiogenesis, genomic instability and mutation, resisting apoptosis, deregulation of cellular energetics and sustaining proliferative signaling.

Each hallmark trait might be the consequence of one or more perturbed pathways and each pathway might be associated with more than one trait. Differential activation of biological pathways and consequent pathway reprogramming, changes in their mutual regulatory feedbacks and summative effect aimed at achieving the hallmark capabilities – are represented in the phenotypic differences of cancers. Among the cancer pathways, some are initiators whereas others are involved as a secondary event. Some pathways are commonly active in cancer conditions, which means they carry minimal diagnostic value. Diagnostics based on these pathways may help reducing the within-class heterogeneity in presently known broader molecular classes. Besides involvement of pathways in cancer may not be static but is rather a dynamic and continuous process. Therefore pathway-based diagnostic profiling of cancer is more useful for the subsequent clinical decision-making, compared to other forms of classifications. For the pathway-based diagnostics in cancer, biomarkers that can uniquely

represent the dysregulated pathway and the corresponding cancer subclass – needs to be explored.

The core of pathway-based marker-search is formed by the following questions: Which pathways significantly influence the outcome and overall phenotype? Which pathways are commonly perturbed/active in more than a single subclass of cancer? Which pathways can uniquely underlie a specific molecular subclass? Which pathways are cancer-initiating pathways and which ones are secondarily activated pathways?

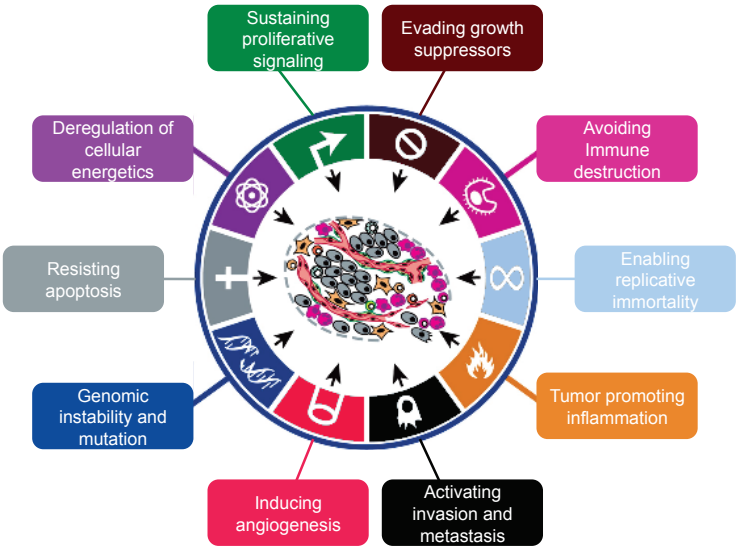


Figure 3: Hallmark processes of cancer
Common hallmark processes drive the initiation and progress of cancer. (Figure source: [75])

1.8.2 Overview about pathway analysis approaches for genomic data

There has been progress in developing the computational prediction algorithms for addressing some of these questions. The goal of such algorithms is typically to identify the sets of pathways differentially perturbed in a given pair of conditions and infer genes that contribute

to the pathway deregulation. The priority of pathway-based diagnostics is to develop the tools that facilitate unsupervised analysis of cancer datasets and thereby can categorize cancers according to the pathway deregulation. Some of the pioneering publications [76-79] have outlined the statistical approaches and associated methodological issues.

The overview and classification of the available algorithms for pathway analysis is shown in *figure 4*.

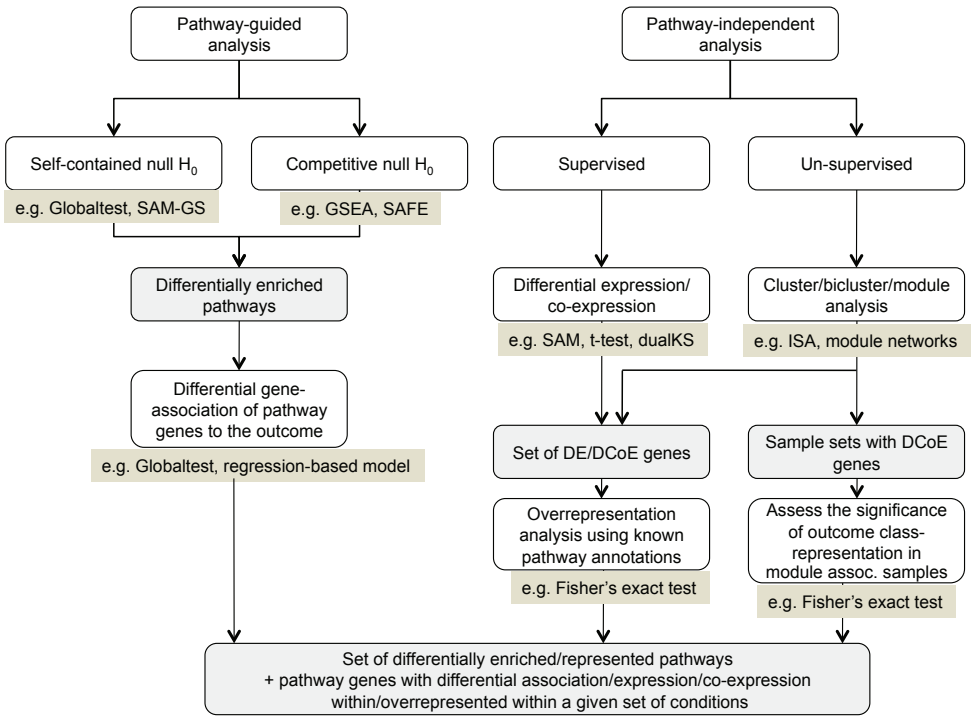


Figure 4: Overview of the pathway-based approaches.
Pipeline with variety of previously published approaches used for the pathway analysis and identifying the pathway marker genes associated with the known condition or response variable

Among the approaches that are pathway-guided and class-supervised, the analysis can be performed by either self-contained or competitive null hypothesis. Self-contained null hypothesis assumes that no gene in the pathway is differentially enriched, whereas the competitive null hypothesis assumes that genes in the given pathway are as differentially

enriched as genes not involved in that pathway. Because the number of genes in pathway is usually too small in comparison to the total number of genes excluding genes in the pathway, the likelihood of null hypothesis rejection is higher in case of self-contained null hypothesis compared to the competitive null hypothesis and thus yields more power [77]. SAM-GS [80](a generalization of SAM [81] for individual genes), globaltest [47] and pathway-significance approach described by Tian *et al* [82] are two published algorithms that assume self-contained null. Geneset enrichment analysis (GSEA) [83] and SAFE [84] are the ones that assume competitive null. Globaltest algorithm also provides a possibility to identify the significant genes that contribute to the differential pathway activity. As an alternative, every gene within each significant pathway (identified with any of the abovementioned methods) can be tested for its association to the categorical or continuous outcome by using the logistic and linear regression model, respectively. Since the test is performed for each gene as covariate within each individual pathway, it is possible that a gene might assume significance in more than a single pathway it is involved in, or can be found significant in one but not in the other pathway.

Another approach is a pathway-independent search for individual genes and subsequently performing the pathway-analysis by using overrepresentation tests with each available genesets or pathways. Besides, the analysis can be class-supervised (e.g. SAM [81], moderated t-test [85]) or unsupervised approaches that include clustering, biclustering (e.g. iterative signature analysis [79] , principal component analysis etc.

The functional derangement of genes and pathways in cancer may not be merely an outcome of their markedly altered gene expression patterns, but a combination of subtle to strong and coherent changes in a set of gene expressions, leading to a significant alteration in the overall pathway activity. In this view, diagnostic marker search should be based on a combination of methods that can explore pathway genes having subtle but coherent class-specific expression changes and ones that search for individual genes with significant and strong gene expression changes.

1.8.3 Limitations of pathway-based analysis

It is obvious that subgroups with differences in pathway activity do not necessarily indicate the differences in survival. Therefore pathway analysis approaches that do not merely rely upon the survival data for either deriving or validating the classes- might have the advantage

in better sensitivity in characterization of novel pathway-based classes and/or signatures. Besides, pathway analysis depends upon the available annotations and interactions. Therefore genes with unknown involvement in specific pathways might go unnoticed. The statistical significance of pathways and each particular gene within the pathway might be influenced by the size (number of genes involved in the pathway) of the pathway. Pathways are defined based on the canonical functional role reported by published literature sources. However many interactions are context-specific and stromal interactions may alter the canonically known interaction. Therefore it could be a chance that the assumption of similar activity of genes *in vivo* as reported by the available experimental evidence holds true.

1.9 Implications of pathway-based diagnostics on breast cancer therapeutics

Historically, the improved diagnostics has provided opportunities for more informed therapeutics. Since the first ever isolation of estrogen receptor[86] from breast tumors and introduction of mammography in 1967, the diagnostics has improved considerably with the advances in laboratory methods and scanning techniques. Improvements radiotherapy and surgery have also contributed to the quality improvements in multimodal management of breast cancer. Since past decade, molecular research has paved the way for improvement in diagnostics and therapeutics.

The hallmarks concept [75] of cancer implicates the focus of novel treatment strategies at targeting the specific key biological pathways that underlie one or more of the hallmarks. The genes that are computationally identified as significant within each differentially active pathways within a given cancer class - are the potential therapeutic targets. Computational approaches, similar to the ones described in the previous section – provides an opportunity to identify the novel markers of diagnostics and therapeutics.

Ideally the therapeutic intervention should selectively target the tumor driver pathways and reversibly rectify the alterations of molecular pathways with minimum possible effect on non-cancer cells and with minimum possible systemic adverse effects. However, currently there is no such treatment option available and therefore clinicians make the decisions by weighing benefit against the risk of adverse effects. Gene or pathway based therapy options mainly target a specific pathway(s) and therefore might confer lower systemic risk of adverse effects.

Because the observed effect on pathway genes and possible effect of targeting the gene within the significant pathway are the consequence of complex multidimensional interactions between proteins and genes, it requires further systems biology work-up to model the network and simulate the effect of perturbations in the network at different genes and select the best possible target(s). The process of drug discovery, design and clinical trials takes a long time and involves a considerably high cost without any certainty. Despite of these difficulties, therapeutics based on molecular diagnostics and pathways has become a priority.

1.9.1 Limitations of currently available chemotherapy options

The following are the major limitations of the currently available chemotherapy options:

1. Most currently available chemotherapeutic options are associated with cytotoxicity, which is not limited to cancer cells. The cytotoxic effect varies depending upon the dose and administration schedule. Such as – cytotoxic effect of anthracycline can cause cardiac toxicity by damage to the cardiomyocytes, bone marrow suppression, etc.

2. Currently chemotherapy is administered in triple negative patients, in Her2+ patients combined with Trastuzumab and in some of the high-risk categorized ER+/Her2– patients. However, complete response is observed only in a fraction of patients receiving the chemotherapy. Previous trials showed that pathologic complete response to pre-operative chemotherapy in hormone-negative breast cancer varies between 9% to 26% [87].

3. While the combined regimens (polychemotherapy) might help in improving the response, it also increases the side effects.

4. Available chemotherapy options may provide significant though relatively shorter duration of survival benefit in metastatic breast cancer. Median survival with first line chemotherapy is up to 25 months [88] and even shorter for second or further lines of chemotherapy.

5. The criteria of decision-making for chemotherapy are limited. Considerable fraction of patients who might have responded to hormonal therapy only, receive unnecessary chemotherapy.

1.9.2 Advantages of pathway-based therapy compared to chemotherapy

The main goal of pathway-targeted therapy is to target the derangement of specific pathways and thereby addressing one or more hallmark characteristics of cancer. For example, pathway targeted therapy by VEGF signaling inhibitors may target the enhanced angiogenesis, proliferative signaling, invasion and metastatic properties of cancer cells. Successful therapy should not only improve overall patient survival and arrest the progression of cancer but should also maintain the quality of life by minimizing the treatment adverse effects.

Figure 5 shows the overview of currently available strategies aimed at specific hallmark characteristics of cancer [75].

The basis of pathway-based therapy is in successful diagnosis of pathway derangement. Once the aberrant key pathway is known, suitable target within that pathway is identified. The following are the benefits of a well-planned pathway-based therapy:

1. Combination of pathway-based therapy together with available chemo- or endocrine therapy can reduce the chances of resistance and recurrence by achieving early response and preventing secondary involvement of more biological pathways.
2. Monodrug therapy can be sufficient in case the suitable target is determined and that the therapy targets the most dominant pathway that is responsible for the cancer growth.
3. Most pathway-based drugs are cytostatic. Therefore, the effect is more likely to be reversible.
4. Many of the drugs are administered orally, in contrast to chemotherapy where large majority of drugs have to be administered intravenously.
5. While pathway-based therapy might also cause systemic side effects just as do chemotherapy, it is anticipated that novel drugs with fewer side effects and optimized targeting strategy would be able to reduce the burden of side effects and thereby would relieve morbidity.

In the *figure 5*, the treatment strategies for which the drugs are either under design, development or trial are shown in orange color fonts. Drugs that are already approved for breast cancer are shown in green fonts. This figure shows that large majority of therapeutic options have not yet entered into practice.

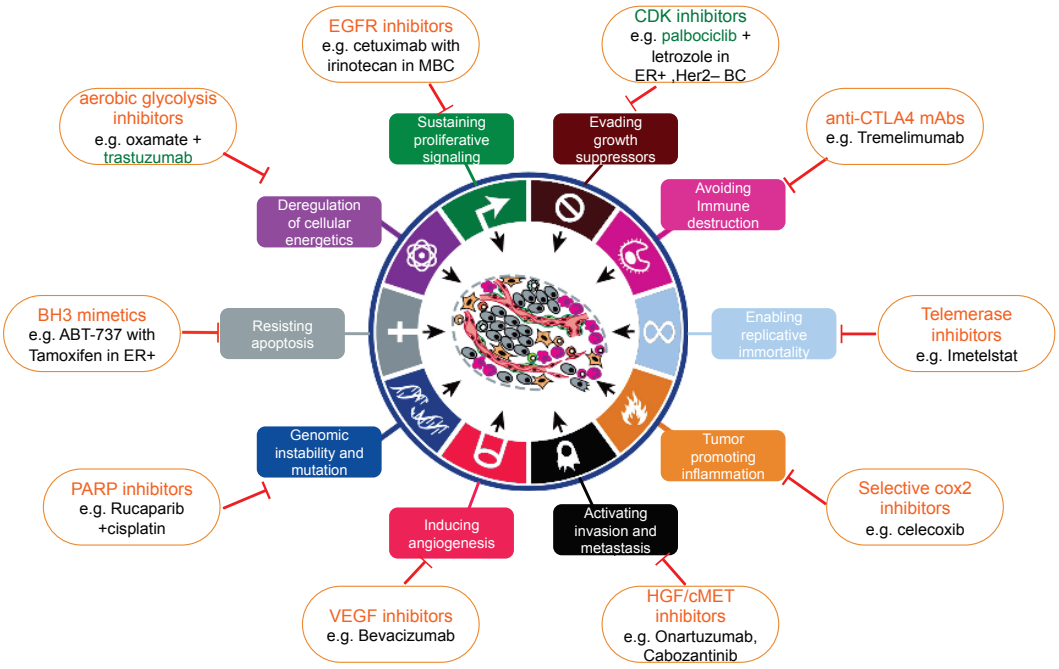


Figure 5: Hallmark-based illustration of novel targeted therapeutic strategies
Novel treatment strategies of cancer target the specific hallmarks of cancer by targeting the specific dysregulated pathways or selected molecule within in the pathway. Orange fonts indicate strategies under trial and green fonts indicate approved strategies or drugs for treatment of breast cancer. (Figure adapted from [75])

1.9.3 Challenges in pathway-guided diagnostics

The term *prognostic factor* is used to define any measurement available at the time of diagnosis or surgery that is associated with clinical outcome in the absence of systemic adjuvant therapy. On the other hand, the term *predictive factor* defines any measurement associated with response or non-response to therapy. Substantial number of studies has focused on prognostication but relatively fewer studies have investigated the predictive markers. The priority in interest of advancing cancer therapeutics has to be on predictive diagnostics. Precisely estimated activity of key biological pathways and networks may help identify the key biological pathways shaping the phenotype. Classification based on this approach can provide robustness. Biologically, robustness means an ability of pathways and networks to overcome the effect of other biological processes on phenotype. In terms of therapeutics, robustness means that the inferred classes can robustly predict response to

corresponding pathway- or network-targeted therapy. It is difficult to achieve this objective of pathway-based approach in reality.

The following are the challenges in the pathway-guided diagnostics:

Despite of encouraging results from clinical trials, robust and efficient predictive biomarkers are yet to be discovered. It sounds intuitively logical to hypothesize that the best predictive marker of pathway directed therapy is aberrantly expressed or mutated. However, the experience with Trastuzumab and its predictive marker Her-2 suggests that expression of the target could probably be a good marker of biology but not the only response-predictive marker. In reality, only 30% of patients with Her-2 overexpression respond to Trastuzumab [89]. Various explanations are proposed for the resistance, such as compensatory activation of other members of HER family [90] or other signaling pathway, inhibition of trastuzumab activity by fragments of Her-2 extracellular domain cleaved from the Her-2 receptor[91], formation of IGF-IR/HER-2 heterodimer [92].

The differences between the pathway-based classes might be subtle and the priority should be to acquire an ability to develop a set of pathway activity-based predictive markers that guide selecting individualized pathway-guided therapy.

In general, it is anticipated that the levels of markers usually correlate with the response. This assumption appears to be partly true in context of some but not all markers. Such as ER positivity and response to endocrine therapy.

Insufficient sample quantity, improperly localized biopsy site etc could also lead to misleading conclusions.

Within-tumor heterogeneity is one more challenge that is difficult to be accounted for in most high throughput analyses. Therefore the signatures and target selection made by these studies could overestimate the size of possible responders.

Despite of the improved efficiency and lowered cost of target identification by computational analysis of high-throughput omics data, high-throughput screening and computation drug design methods, the cost per novel drug approved has increased substantially [93] with no significant reduction in time to introduce the drug in practice. This presents as a barrier to efforts of innovative novel drug discovery and development.

2 Aims of the thesis

Improving the understanding about molecular heterogeneity is essential step towards defining the molecular classification of breast cancer that has translational value. Existing definitions of molecular portraits are based on unsupervised methods on high-throughput omics data and differences of survival according to the classes are demonstrated.

The study involved understanding the known molecular classification, predict the relevant potential transcriptional mechanisms and then to identify novel classes having diagnostic, therapeutic and prognostic significance based on the status of key transcription factors, deriving the novel class-specific signature based on pathway-based approach, recognizing the interaction of key signaling pathways in defined classes. This approach helps understanding the advantages and limitations of the existing breast cancer molecular classification by focus on pathways and processes that have considerably higher prognostic impact rather than merely focusing on individual genes. This work might help contributing an additional perspective in understanding the tumor driver mechanisms by the application of the pathway-/process based approach, in contrast to large volume of studies that are based on individual gene-based approaches. The study started with the investigation of the transcription factors involved in the regulatory networks of genes defining the existing molecular portraits. The study also involved understanding the context-specific regulation of mRNAs bytheir potential regulator miRNAs.

The following are the main aims of the study:

1. To study the promoter composition of subtype-distinguishing genes and predict the key transcription factors regulating these genes and thus having potential functional role in the phenotypic diversity of subtypes.

2. To identify the master regulator transcription factor with functional relevance to the subtypes

3. To identify the differentially activated pathways with respect to the categorization based on the status of one such a master regulator (p53).

4. To identify the class-specific candidate marker genes that influence the differential activity of pathways with respect to the status of one master regulator transcription factor (here p53 mutation status).

4. To evaluate the signature genes for their prognostic significance by controlling for the existing determinants of patient survival.

5. To elucidate the context-specific potential regulatory miRNAs-mRNA modules in breast cancer expression profiles with reference to the newly identified molecular classes of prognostic significance.

3 Materials

3.1 Breast cancer expression profiles

The project was facilitated by the public access to three datasets from Norwegian and Swedish cohorts of breast cancer. The MicMa dataset was the primary or learning dataset. Data from two other cohorts – Uppsala (N=251) and Ullevål (N=76) were used as test datasets. Table 3 provides the overview of the datasets.

Table 3: Overview of the datasets used in this study

Dataset	Geographic profile	Sample profile	Years of study	Type of data	#Samples used	Platform	Source
MicMa	Norwegian	Primary human breast cancer	1995-1998	mRNA expressions	114	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	GSE19783
				miRNA expressions	100	Agilent-019118 Human miRNA Microarray 2.0 G4470B	GSE19783
Uppsala	Swedish	Primary human breast cancer	1987-1989	mRNA expressions	251	Affymetrix Human Genome U133A and U133B arrays	GSE3494
Ullevål	Norwegian	Primary human breast cancer	1990-1994	mRNA expressions	76	42 K cDNA microarrays	Stanford Microarray database

3.1.1 MicMa dataset

Out of the 900 patients diagnosed with breast cancer diagnosed between May 1995 and December 1998 at Oslo, mRNA expression profiles are available for 115 samples. After performing the quality control and clinical data availability, 111 samples were included in the analysis for this project. Patients less than 55 years with grade 2-3 and/or nodal involvement were treated with CMF chemotherapy and additional Tamoxifen if hormone receptor positive.

Hormone receptor positive patients who were older than 55 years received only Tamoxifen. Hormone receptor negative cases with grade 2-3 or nodal involvement in the 55-65 years age group received CMF regimen but older patients were not administered adjuvant therapy [94]. Complete clinical data is available in the respective publications. Follow up time was about 10 years in this cohort.

3.1.2 Uppsala Dataset

Out of the 315 diagnosed primary breast cancer patients registered between January 1, 1987 to December 31, 1989 [95] in the Uppsala county of Sweden, a subset of 251 samples was processed for TP53 mutations and microarray data. Therefore this subset has been used for the analysis. Systemic adjuvant therapy was administered to all patients with nodal involvement. Premenopausal women received chemotherapy and postmenopausal women received endocrine treatment. About 55% patients did not receive adjuvant therapy. The median follow-up duration was 122 months [96].

3.1.3 Ulleval dataset

This dataset consist of 80 samples (76 included after quality control) out of a series of 212 primary breast cancer cases collected at the Ullevål Hospital between 1990 and 1994. Large fraction of patients with larger tumor size is included in this dataset. Follow up period for this dataset was about 12 to 16 years. The group received chemo according to existent management guidelines published by the Norwegian Cancer society.

3.2 Microarray technology platforms

3.2.1 Human whole genome oligoarray (Agilent)

The breast cancer samples from MicMa cohort (N=111) are based on the Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe name version) –GEO accession: GPL6480 platform. This platform provides a possibility to hybridize the RNA transcripts to ~41,000 unique 60-mer probes. Agilent provides the feature extraction tools and the annotation files with mapping of probes to the genomic transcripts.

3.2.2 Human genome U133 oligoarray (Affymetrix)

Expression profiles from the Swedish Uppsala cohort are based on the Affymetrix Human Genome U133 (hgu133) platform. The hgu133 platform consists of two arrays – hgu133A and hgu133B, together consist of 44928 probe sets that represent >39,000 transcripts derived from ≈33,000 human genes. Probesets in the hgu133A represents RefSeq database sequences and probe sets related to sequences previously represented on the Human Genome U95Av2 Array. Probesets of hgu133B represents EST clusters.

3.2.3 Human genome cDNA arrays

Expression profiles for the Ulleval dataset are based on the cDNA arrays. The protocol uses amplified RNA from the tumor material. The platform provides cDNA microarray chip with more than 42000 elements. Full details of this platform are accessible from the Stanford Microarray database (http://smd.princeton.edu//)

3.2.4 Human miRNA Microarray (Agilent)

MiRNA expression profiles from the MicMa were based on the Human miRNA Microarray 2.0 G4470B (Agilent). This platform consists of 723 human and 76 human viral microRNAs from the Sanger database v.10.1. Full annotations are available at GEO accession : GPL8227.

3.3 TP53 mutation data

MicMa and Ullevål dataset uses TP53 mutations data derived from tumor DNA by prescreening exons 2–11 with temporal temperature gradient gel electrophoresis (TTGE) protocol. Whereas TP53 mutation data from Uppsala cohort was based on solid phase sequencing on p53 amplified from tumor cDNA using PCR [95] and analysis was performed on exons 2-11.

4 Summary of papers

Paper I: Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes.

***BMC Genomics*, 2012; 13:199.**

This piece of work involves an approach of exploration of transcriptional regulatory networks that underlie the heterogeneity of breast cancer. Regulation of gene regulation in eukaryotes is highly complex and depends on sets of transcription factors rather than individual transcriptional factors. In normal or diseased human tissues, functional diversity is achieved by the combination of a small number of transcription factors whose activities are modulated by diverse sets of conditions. Sets of common transcription factors might be responsible for activation or repression of sets of target genes that act in coherent manner to produce the phenotype. This analysis is based on the hypothesis that overrepresented transcriptional factor binding site motifs within a group of co-expressed gene promoter sequences are more likely to be co-regulated by a set of transcription factors and can have a role in transcriptional activation. We aimed to identify distinct promoter composition and overrepresentation of key transcription factors in a set of co-expressed genes that give rise to the breast cancer subtype-specific expression patterns. We have applied a pipeline that includes transcription factor binding site overrepresentation analysis of putative promoter regions of the genes for distinguishing between five molecular subtypes. The transcription factor genes were mapped based on the overrepresented transcription factors in order to validate this hypothesis in real mRNA expressions within each predicted subtype. In order to pursue this analysis, mRNA expression profiles of breast cancer from previously published dataset of Norwegian cohort consisting of 111 samples were first categorized into molecular subtypes by using PAM50 classifier algorithm. In the classified samples, the actual mRNA expression values of transcription factor genes were correlated with fold-factor overrepresentation of the corresponding transcription factor binding site motif.

Approaches as the one used in this paper demonstrate the differential overrepresentation of transcription factors binding sites corresponding to the differential expression pattern of genes

referred as molecular subtypes. Besides, the transcription factors corresponding to the significantly overrepresented transcription factor binding sites are representative of the biological characteristics of the subtypes. This work implicates further experimental studies to investigate and validate the precise regulatory interactions.

Paper II: Potential tumorigenic programs associated with TP53 mutation status reveal role of VEGF pathway

***British Journal of Cancer*, 2012; 107:10(1722-1728)**

Targeting differentially activated or perturbed tumor pathways is the key idea in individualized cancer therapy, which is emerging as an important option in treating cancers with poor prognostic profiles. With the view of obtaining better understanding about the molecular heterogeneity and for enhancing the translational potential of molecular classes, it is essential to identify novel classes that have prognostic, therapeutic and diagnostic significance. Known prognostic markers-based classification not only provide the insight about biological pathway activity differences between different phenotypes but also provides an opportunity of exploring other associated significant markers and thereby in help creating novel diagnostically meaningful classification. TP53 mutation status is known as a core determinant of survival in breast cancer. The pathways disrupted in association with TP53 mutation status in tumors are not well characterized.

TP53 is a key regulatory gene and an independent predictor of clinical progression, prognosis and therapeutic response of breast cancers, apart from the molecular subtypes. However, the driver pathways underlying the differential phenotype and their underlying regulatory interactions remain to be elucidated. TP53 mutation type (e.g. missense, frameshift, splice and nonsense mutations) and its location (such as within the CpG island, DNA-binding region and location in terms of the domain) of TP53 gene, subsequent influence on the extent of loss of transactivation ability or structural alteration of TP53 determines the prognosis and survival of breast cancers.

In this study, we stratify breast cancers based on their *TP53* mutation status and identify the set of dysregulated tumorigenic pathways and corresponding candidate driver genes using

breast cancer gene expression profiles. Expressions of these genes were evaluated for their effect on patient survival first in univariate models, followed by multivariate models with TP53 status as a covariate.

The most strongly differentially enriched pathways between breast cancers stratified by *TP53* mutation status include in addition to TP53 signaling, several known cancer pathways involved in renal, prostate, pancreatic, colorectal, lung and other cancers, and signaling pathways such as calcium signaling, MAPK, ERBB and vascular endothelial growth factor (VEGF) signaling pathways. We found that mutant TP53 in conjunction with active estrogen receptor (ER) signaling significantly influence survival. We also found that upregulation of *VEGFA* mRNA levels in association with active ER signaling is a significant marker for poor survival, even in the presence of wild-type TP53.

Consistent with the survival differences, we identified the class-specific candidate marker genes in each group. In contrast to the routinely used methods focusing on differential expressions of genes, we successfully applied a combined strategy that involved methods accounting for the condition-specific association by coherent expression or biased expression of genes. Genes driving the abnormal pathway activity were identified.

This work concluded that mutation status of *TP53* in breast cancer involves wide ranging derangement of signaling, metabolic and other pathways. Among the candidate genes of the significantly deranged pathways, *VEGFA* expression status is an important marker of survival even when controlled by *TP53* mutation status. Interestingly, independent of the *TP53* mutation status, the survival effect of *VEGFA* was found significant in patients with active ER signaling (ER/PgR+), but not in those with ER/PgR− status. Therefore, this work proposes more studies to focus on the role of complex interplay between TP53, ER and VEGF signaling from therapeutic and prognostic context in breast cancer.

Paper III: Implications of *VEGFA* upregulation on microRNA-mRNA Modules in Breast Cancers

Interactions between MicroRNAs (miRNAs) and mRNAs form the crucial components of post-transcriptional regulation of gene expression both in healthy as well as in the malignant

state of the tissues. Given the tissue- and context-specificity of their function, it is useful to decipher modules of miRNAs and their targets that exhibit specific functionally correlated expression patterns in previously known classes of cancers. Many of these classes were earlier studied only by using their mRNA expressions and then the disease-specific networks were predicted based on protein-protein interactions.

Activation of vascular endothelial growth factor (VEGF) pathway in breast cancer has been associated with high microvascular density, influencing prognosis and response to conventional hormonal therapy. In this study, breast cancers are categorized into a subgroup with upregulated VEGFA mRNA and a subgroup with normal/downregulated VEGFA mRNA levels. By using previously published miRNA and mRNA expression dataset of a Norwegian cohort from the same breast cancer cases, differential correlative expression patterns of miRNA modules and their predicted targets that overlap differentially expressed genes between the two groups of breast cancers are studied.

Differential expression analysis revealed 36 miRNAs and 162 gene features differentially expressed between the two VEGFA expression groups. Predicted mRNA targets of miRNAs were obtained from the predicted miRNA target database. For each of differentially expressed miRNA, its correlation with the mRNA expression of its corresponding predicted target genes was computed. Among the profiles with VEGFA-upregulation, miR-18a/18b/19a cluster, miR-9/9*, miR-342-3p/5p etc and downregulation of miR-149, miR-135b, miR-449a was observed. Anticorrelated targets of miRNAs upregulated in VEGFA+ group were enriched for angiogenesis pathway, biological processes of vasculature development and TGFβ signaling and focal adhesion. Whereas the anti-correlated targets of downregulated miRNAs in VEGFA+ group were found significantly associated with EGFR pathway.

In this study, differential anti-correlative patterns between miRNAs and their targets with respect to the VEGFA expression status are found. More work is proposed for the validation of the findings on an independent dataset.

5 Discussion

With the evolution of high-throughput technologies, omics profiling has generated large-scale data. This has implicated the need of developing efficient computational approaches, tools and methodological pipelines for genomic data analysis and correct interpretation. DNA microarrays remains to be a technology with proven track record in providing the transcriptomic snapshot. Molecular breast cancer research has benefited considerably during the past decade as a result of these technologies.

Pathway-based genomics and the study of the cancer heterogeneity require large size of cohorts where the independent pathway-based alterations can be characterized with sufficient statistical power. This implies the need of genomic data from large cohorts. With the exception of a few, there has been an increasing trend towards unrestricted public accessibility of data among the experimental biologists and cancer labs. As a result, large number of expression profile datasets and clinical profiles of patients have become available in public repositories. This has made it possible for the independent researchers to combine the datasets from different cohorts to achieve sufficient statistical power for the pursuit of biomarker discovery. Besides projects such as TCGA are going to be helpful to bioinformaticians and medical scientists in pursuing freelance research that can promote innovation.

5.1 Methodological considerations

This study involves the publicly available data from the Swedish and Norwegian cohorts. The analysis was performed on the raw data downloaded from the respective sources. Preprocessing was then performed with quality control and elimination of samples with either poor quality or high fraction of missing values.

Paper I and II used PAM50-centroid based method [97] for categorizing the molecular portraits in expression profiles, rather than using the previously used centroid-based approach [98]. PAM50-centroid based method is based on small signature size and therefore improves objectivity. Besides it has shown higher predictive value for complete response among the patients with chemotherapy.

Paper II uses two different approaches – pathway-based and gene-based methods for identifying the p53 mutation class-specific signatures of breast cancer. This combined approach enhances the sensitivity of marker search, as it does not use strict cut-offs of differential expression methods but identifies genes with weaker differential expression active within the pathway. Thus it helps deriving more biological information compared to the previously published signatures.

Intra-tumor heterogeneity is one of the significant limitations of the work presented in this thesis. Concurrent occurrence of cell subpopulations with differing clonality within a single tumor is referred as intra-tumoral heterogeneity. These cell subpopulations possess different sets of genomic alterations. Even though tumor cells are believed to have originated from the same progenitor cell, during the evolution of the tumor cells are believed to achieve the diversity in genomic alterations, where these alterations confer differing degree and types of hallmark characteristics to the cell subpopulations. Some tumors show dominance of one clonal subpopulation with stable chromosomal structure (monogenomic), whereas others show presence of multiple clonal subpopulations at one or more locations (polygenomic) [99]. Tumor progression, aggressiveness, biological characteristics and even therapeutic response-can have considerable influence from the degree of intra-tumoral heterogeneity. Intra-tumoral heterogeneity poses to be an issue for interpretation of microarray-based expression profiles because there is no way to ascertain which clonal subpopulations is represented in the biopsy material and subsequently derived expression profile. Among the major implications of this unaddressed source of intra-tumor heterogeneity within tumor comes mainly from the cancer stem cells, as presence of stem cell population in the tumor might indicate poor response to chemotherapy [100] and increase the likelihood of recurrence. It is necessary to acknowledge that results from this study do not account for the cell population heterogeneity arising from the clonal architecture of the tumor. Despite of this limitation, it has been argued that intra-tumor heterogeneity being a continuous and accumulative process, most subpopulations might represent the fundamental genomic alterations of diagnostic and prognostic significance [101].

Paper I presents statistical overrepresentation of known transcription factor binding families in the promoter sequences of subtype-relevant clusters. The subtype-relevant clusters were not based on significance of co-expression within each subtype. Besides, the statistical overrepresentation as seen in this work only accounts for the overrepresentation of known or

predicted potential TFBS motifs from a proprietary database- Transfac and not uncharacterized motifs or known motifs from other database. However, we believe that Transfac database accommodates most motifs that can be interpreted for possible functional role based on the literature evidences. Previously uncharacterized motifs even when detected, could be difficult to explain their possible functional role.

Even when the cohorts of all three datasets are primary breast cancers, the Ullevål cohort consists of the samples collected from relatively larger tumors and therefore it is likely that these tumors are relatively more advanced compared to the other two cohorts. But still the methodological stratification by TP53 mutations status and ER status might have controlled for any possible bias in the results.

There are between-cohort variations in the treatment protocols of adjuvant regimens and that means that expression patterns of certain genes might vary accordingly, in particular genes influenced by immune response. These differences are not accounted in signatures. However, none of the study focuses on processes with major implications from therapeutic regimen (such as immune response) and therefore the results might not have been significantly biased by the therapeutic differences.

The original aim of paper II was to infer the differentially perturbed pathways by p53 mutation classes (such as missense within DNA-binding region, missense outside DNA-binding region, non-missense etc). However, the analysis remained limited to major classes-wild-type p53 and mutant p53 in breast cancer because of lower number of individual mutation classes and non-availability of large publicly available data on p53 mutations and corresponding expression profiles.

Paper II uses the large cross-platform cohort by merging the Swedish and Ullevål datasets by cross-platform normalization method for the validation of the signatures inferred on the primary dataset. Expression data in these three cohorts is based on different technology platforms. Differences in probe designs, labeling, hybridization, and scanning may lead to the variability in gene expression estimates. Sufficient evidence showing concordance between the cDNA arrays and Agilent or Affymetrix whole genome arrays is lacking. Besides differences in the laboratory protocols, sample collection protocols are another source of variability. However, Affymetrix and cDNA platform-based data were merged using UniGene identifiers in order to compile them as a validateion dataset. Class-specific signatures derived

by analyzing the primary data (based on Agilent whole genome technology) were then validated. There is an evidence that variability of expressions because of the platform differences might not considerably change the model performance[102] in case of classification analysis, which means that sensitivity might not be affected by cross-platform variability. Therefore we consider that sensitivity to find the true positives in inferred signatures would have either remained unchanged or improved by performing analysis on a data from one platform and then validating it on a cross-platform dataset. In addition, it might have also helped eliminating the laboratory and platform-specific bias.

Paper III studies VEGF expression class-specific miRNA-mRNA modular relationship. The two basic assumptions for the study are: 1. Class-specific anti-correlation between the expression values of differentially expressed miRNA and putative target mRNA indicates potential functional regulatory role of the miRNA on the target mRNA. Putative target mRNA means a predicted target site with good mirSVR score (score <= 0.1 obtained by mirSVR algorithm) and conserved miRNA according to microRNA.org [103]. 2. Downregulation or upregulation of mRNA target is a consequence of differentially expressed (in opposite direction) and anti-correlated putative regulatory miRNA and is not as a result of any other factors such as epigenetics, gene-protein and protein-protein interactions etc. These assumptions might not always hold true. MiRNAs act on several pathways and processes and can regulate many genes, however functional role of many miRNAs is not sufficiently proven. Therefore it is could be difficult to prove that the regulatory role of miRNA is the only major role associated with the consequence on the expression of the target. This limitation remains true even in case of experimentally validated regulatory relationship. Many of the regulatory functions of miRNAs are often transient, cell-condition and tissue-specific. Therefore considerable fraction of the inferred miRNA-mRNA regulatory interactions might be false positives. Despite of these limitations, there is no doubt that the computational pipeline used here forms an extremely useful means to formulate a hypothetical miRNA-mRNA regulation network that can be validated by suitable experimental methods.

5.2 Future directions

The work presented in the thesis has got several interesting dimensions. Here some of the possible developments of this work are discussed.

Paper I presents a group of interesting transcription factors that are significantly overrepresented in the subtype-distributing gene promoter sequences. This analysis was performed on proximal promoters (–500 bp to +100 bp from the transcription start site) of the subtype distinguishing genes. The possible directions from this work are the following: 1. Instead of a subset of subtype-classifier genes, genes that follow class-specific significant co-expression can be used in the analysis. 2. This work involves the putative regulatory region of –500 bp to +100 bp relative to the TSS. Because it is known that this region in proximity of the transcription start site has high density of functional transcription factor binding sites. However, there are other regulatory elements that occur in the distal promoter regions and they follow the sequence-based binding. 3. Transcription regulation might often involve a combination of multiple transcription factors binding on a same set of promoters, referred as *cis*-regulatory modules. Linking the *cis*- regulatory modules to a set of class-defining cluster genes could be a possible direction. 4. This analysis searched for the known transcription factor binding sites for the motifs (motif families) included in the Transfac database. Inclusion of motifs from multiple other databases such as Jaspar, ORegAnno after eliminating redundant motifs could be a strategy to expand the spectrum of search. 5. For a given set of transcription factor families that were found significantly overrepresented within the subtype-distinguishing gene promoters, experimental validation using ChIP followed by ChIP-Seq could be performed.

Paper II is presents the pathway analysis of breast cancer expression profiles by using the mutation status of p53 gene (wild-type versus mutant p53 gene). Primary dataset included 111 samples, out of which 73 were included in the wild-type and 38 in the mutant p53 class. It is important to note here that the effect of p53 mutation varies considerably depending upon the location and type of the mutation on the p53 gene and consequent loss of function on resultant p53 protein. Mutation status of p53 can be categorized into subclasses that can be described as TP53 mutation effect groups[104], because of their differing degree of impact on patient survival. Besides, mutations located on specific positions on p53 central DNA binding region are relatively more frequent. Study focusing on the effect of the individual p53 mutation types and mutation effect group could provide interesting insight into the effects of p53 mutation. The plan is to obtain a large publicly available dataset and then to apply the similar methodological pipeline using the individual mutations as well as p53 mutation effect groups (such as missense within DNA-binding region, missense outside DNA-binding region, non-

missense etc) as classes. Such an analysis would reveal a set of biological pathways that differentially activated according to p53 mutation effect group.

Analysis presented in Paper III compares the mRNA and miRNA expression profiles of same breast cancer samples categorized into two classes according to *VEGF* expression status and identified differentially expressed genes and miRNAs. By using a predicted target site database for humans (August 2010 release, available from microRNA.org) having good mirSVR score (score <= 0.1 obtained by mirSVR algorithm) and conserved miRNA, sets of predicted mRNA targets were obtained for each of the differentially expressed miRNAs. Correlation was then computed between class-specific expression values of each of the miRNA and its potential target mRNA. Significantly anti-correlated and differentially expressed (in opposite direction) mRNAs were considered as potential targets. In this analysis pipeline, some modifications are possible. Instead of using the predicted target site database, it is possible to use miRNA target predicting algorithm alone or in combination of one or more of the target site database. MiRNA annotation and their functional GO terms are poorly defined, but one can perform the functional representation of analysis by using custom-made miRNA functional database. The results obtained in this analysis will be validated with a publicly available independent dataset of miRNA and mRNA expression profiles. Moreover, experimental demonstration of miRNA-mRNA target functional interaction in cell-lines with *VEGF* upregulation is possible.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM: **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008**. *Int J Cancer* 2010, **127**(12):2893-2917.
2. Howlader N NA, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). **SEER Cancer Statistics Review, 1975-2010**. In. Bethesda, MD: National Cancer Institute; 2013.
3. Straver ME, Glas AM, Hannemann J, Wesseling J, van de Vijver MJ, Rutgers EJ, Vrancken Peeters MJ, van Tinteren H, Van't Veer LJ, Rodenhuis S: **The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer**. *Breast Cancer Res Treat* 2010, **119**(3):551-558.
4. Mook S, Schmidt MK, Viale G, Pruneri G, Eekhout I, Floore A, Glas AM, Bogaerts J, Cardoso F, Piccart-Gebhart MJ *et al*: **The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study**. *Breast Cancer Res Treat* 2009, **116**(2):295-302.
5. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ: **Clinical application of the 70-gene profile: the MINDACT trial**. *J Clin Oncol* 2008, **26**(5):729-735.
6. Sparano JA: **TAILORx: trial assigning individualized options for treatment (Rx)**. *Clinical breast cancer* 2006, **7**(4):347-350.
7. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, Cheang MC, Gelmon K, Nielsen TO, Blomqvist C *et al*: **Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies**. *PLoS Med* 2010, **7**(5):e1000279.
8. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, Bakx N, Lahti-Domenici JS, Bruinsma TJ, Warmoes MO, Bernards R *et al*: **Converting a breast cancer microarray signature into a high-throughput diagnostic test**. *BMC Genomics* 2006, **7**:278.
9. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T *et al*: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer**. *N Engl J Med* 2004, **351**(27):2817-2826.
10. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J *et al*: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer**. *Lancet* 2005, **365**(9460):671-679.
11. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF: **The prognostic role of a gene signature from tumorigenic breast-cancer cells**. *N Engl J Med* 2007, **356**(3):217-226.
12. Heller T, Kirchheiner J, Armstrong VW, Luthe H, Tzvetkov M, Brockmoller J, Oellerich M: **AmpliChip CYP450 GeneChip: a new gene chip that allows rapid and accurate CYP2D6 genotyping**. *Ther Drug Monit* 2006, **28**(5):673-677.
13. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, Lecocke M, Metivier J, Booser D, Ibrahim N *et al*: **Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer**. *J Clin Oncol* 2004, **22**(12):2284-2293.

14. Rouzier R, Pusztai L, Delaloge S, Gonzalez-Angulo AM, Andre F, Hess KR, Buzdar AU, Garbay JR, Spielmann M, Mathieu MC *et al*: **Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer**. *J Clin Oncol* 2005, **23**(33):8331-8339.
15. 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**. *Proc Natl Acad Sci U S A* 2005, **102**(10):3738-3743.
16. Tutt A, Wang A, Rowland C, Gillett C, Lau K, Chew K, Dai H, Kwok S, Ryder K, Shu H *et al*: **Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature**. *BMC Cancer* 2008, **8**:339.
17. Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, Nelson E, Mone M, Hansen H, Buys SS, Rasmussen K *et al*: **Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay**. *Breast Cancer Res* 2006, **8**(2):R23.
18. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT *et al*: **A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen**. *Cancer Cell* 2004, **5**(6):607-616.
19. Davis LM, Harris C, Tang L, Doherty P, Hraber P, Sakai Y, Bocklage T, Doeden K, Hall B, Alsobrook J *et al*: **Amplification patterns of three genomic regions predict distant recurrence in breast carcinoma**. *The Journal of molecular diagnostics : JMD* 2007, **9**(3):327-336.
20. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L *et al*: **Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma**. *Clin Cancer Res* 2004, **10**(16):5367-5374.
21. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S *et al*: **Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study**. *JAMA* 2006, **295**(21):2492-2502.
22. Spitale A, Mazzola P, Soldini D, Mazzucchelli L, Bordoni A: **Breast cancer classification according to immunohistochemical markers: clinicopathologic features and short-term survival analysis in a population-based study from the South of Switzerland**. *Ann Oncol* 2009, **20**(4):628-635.
23. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, Watson M, Davies S, Bernard PS, Parker JS *et al*: **Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer**. *J Natl Cancer Inst* 2009, **101**(10):736-750.
24. Marcus JN, Watson P, Page DL, Narod SA, Lenoir GM, Tonin P, Linder-Stephenson L, Salerno G, Conway TA, Lynch HT: **Hereditary breast cancer: pathobiology, prognosis, and BRCA1 and BRCA2 gene linkage**. *Cancer* 1996, **77**(4):697-709.
25. Vincent-Salomon A, Gruel N, Lucchesi C, MacGrogan G, Dendale R, Sigal-Zafrani B, Longy M, Raynal V, Pierron G, de Mascarel I *et al*: **Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity**. *Breast Cancer Res* 2007, **9**(2).
26. Walsh MM, Bleiweiss IJ: **Invasive micropapillary carcinoma of the breast: eighty cases of an underrecognized entity**. *Hum Pathol* 2001, **32**(6):583-589.
27. Nassar H, Wallis T, Andea A, Dey J, Adsay V, Visscher D: **Clinicopathologic analysis of invasive micropapillary differentiation in breast carcinoma**. *Mod Pathol* 2001, **14**(9):836-841.

28. Papadatos G, Rangan AM, Psarianos T, Ung O, Taylor R, Boyages J: **Probability of axillary node involvement in patients with tubular carcinoma of the breast**. *Br J Surg* 2001, **88**(6):860-864.
29. Shousha S, Coady AT, Stamp T, James KR, Alaghband-Zadeh J: **Oestrogen receptors in mucinous carcinoma of the breast: an immunohistological study using paraffin wax sections**. *J Clin Pathol* 1989, **42**(9):902-905.
30. Venable JG, Schwartz AM, Silverberg SG: **Infiltrating cribriform carcinoma of the breast: a distinctive clinicopathologic entity**. *Hum Pathol* 1990, **21**(3):333-338.
31. Sastre-Garau X, Jouve M, Asselain B, Vincent-Salomon A, Beuzeboc P, Dorval T, Durand JC, Fourquet A, Pouillart P: **Infiltrating lobular carcinoma of the breast. Clinicopathologic analysis of 975 cases with reference to data on conservative therapy and metastatic patterns**. *Cancer* 1996, **77**(1):113-120.
32. Marchetti A, Buttitta F, Pellegrini S, Campani D, Diella F, Cecchetti D, Callahan R, Bistocchi M: **p53 mutations and histological type of invasive breast carcinoma**. *Cancer Res* 1993, **53**(19):4665-4669.
33. Soomro S, Shousha S, Taylor P, Shepard HM, Feldmann M: **c-erbB-2 expression in different histological types of invasive breast carcinoma**. *J Clin Pathol* 1991, **44**(3):211-214.
34. Ercan C, van Diest PJ, van der Ende B, Hinrichs J, Bult P, Buerger H, van der Wall E, Derksen PW: **p53 mutations in classic and pleomorphic invasive lobular carcinoma of the breast**. *Cellular oncology* 2012, **35**(2):111-118.
35. Monk N, Efroni S, Schaefer CF, Buetow KH: **Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis**. *PLoS One* 2007, **2**(5):e425.
36. Guo Z, Zhang T, Li X, Wang Q, Xu J, Yu H, Zhu J, Wang H, Wang C, Topol EJ *et al*: **Towards precise classification of cancers based on robust gene functional expression profiles**. *BMC Bioinformatics* 2005, **6**:58.
37. Pritchard KI: **Endocrine therapy of advanced disease: analysis and implications of the existing data**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2003, **9**(1 Pt 2):460S-467S.
38. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
39. Matthews L, D'Eustachio P, Croft D, de Bono B, Gopinath G, Jassal B, Lewis S, Schmidt E, Vastrik I, Wu G *et al*: **An Introduction to the Reactome Knowledgebase of Human Biological Pathways and Processes**. *NCI Nature Pathway Interaction Database* 2007.
40. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Res* 2009, **37**(Database issue):D674-679.
41. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
43. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome Res* 2003, **13**(9):2129-2141.

44. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J *et al*: **The BioPAX community standard for pathway data sharing**. *Nat Biotechnol* 2010, **28**(9):935-942.
45. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A *et al*: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**(4):524-531.
46. Balmain A: **Cancer as a complex genetic trait: Tumor susceptibility in humans and mouse models**. *Cell* 2002, **108**(2):145-152.
47. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics (Oxford, England)* 2004, **20**(1):93-99.
48. Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification**. *PLoS Comp Biol* 2008, **4**(11):e1000217.
49. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG: **Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways**. *Nat Genet* 2006, **38**(8):896-903.
50. Norberg T, Klaar S, K\"arf G, Nordgren H, Holmberg L, Bergh J: **Increased p53 mutation frequency during tumor progression--results from a breast cancer cohort.** *Cancer Res* 2001, **61**(22):8317--8321.
51. D'Assoro AB, Leontovich A, Amato A, Ayers-Ringler JR, Quatraro C, Hafner K, Jenkins RB, Libra M, Ingle J, Stivala F *et al*: **Abrogation of p53 function leads to metastatic transcriptome networks that typify tumor progression in human breast cancer xenografts.** *Int J Oncol* 2010, **37**:1167-1176.
52. Aas T, Børresen AL, Geisler S, Smith-Sørensen B, Johnsen H, Varhaug JE, Akslen LA, Lønning PE: **Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients.** *Nat Med* 1996, **2**:811-814.
53. Bertheau P, Turpin E, Rickman DS, Espié M, De Reyniès A, Feugeas J-P, Plassa L-F, Soliman H, Varna M, De Roquancourt A *et al*: **Exquisite Sensitivity of TP53 Mutant and Basal Breast Cancers to a Dose-Dense Epirubicin−Cyclophosphamide Regimen**. *PLoS Med* 2007, **4**:10.
54. Takahashi S, Moriya T, Ishida T, Shibata H, Sasano H, Ohuchi N, Ishioka C: **Prediction of breast cancer prognosis by gene expression profile of TP53 status.** *Cancer Sci* 2008, **99**:324-332.
55. Berns EM, Foekens JA, Vossen R, Look MP, Devilee P, Henzen-Logmans SC, Van Staveren IL, Van Putten WL, Inganas M, Meijer-Van Gelder ME *et al*: **Complete sequencing of TP53 predicts poor response to systemic therapy of advanced breast cancer.** *Cancer Res* 2000, **60**:2155-2162.
56. Molchadsky A, Rivlin N, Brosh R, Rotter V, Sarig R: **p53 is balancing development, differentiation and de-differentiation to assure cancer prevention**. *Carcinogenesis* 2010, **31**(9):1501-1508.
57. Teodoro JG, Evans SK, Green MR: **Inhibition of tumor angiogenesis by p53: a new role for the guardian of the genome**. *J Mol Med (Berl)* 2007, **85**(11):1175-1186.
58. Lebedeva MA, Eaton JS, Shadel GS: **Loss of p53 causes mitochondrial DNA depletion and altered mitochondrial reactive oxygen species homeostasis**. *Biochim Biophys Acta* 2009, **1787**(5):328-334.
59. Kawauchi K, Araki K, Tobiume K, Tanaka N: **P53 regulates glucose metabolism through an IKK-NF-kappa B pathway and inhibits cell transformation**. *Nat Cell Biol* 2008, **10**(5):611-618.

60. Suzuki S, Tanaka T, Poyurovsky MV, Nagano H, Mayama T, Ohkubo S, Lokshin M, Hosokawa H, Nakayama T, Suzuki Y *et al*: **Phosphate-activated glutaminase (GLS2), a p53-inducible regulator of glutamine metabolism and reactive oxygen species**. *Proc Natl Acad Sci U S A* 2010, **107**(16):7461-7466.
61. Budanou AV, Lee JH, Karin M: **Stressin' Sestrins take an aging fight**. *Embo Mol Med* 2010, **2**(10):388-400.
62. Bailey ST, Shin H, Westerling T, Liu XS, Brown M: **Estrogen receptor prevents p53-dependent apoptosis in breast cancer**. *Proc Natl Acad Sci U S A* 2012, **109**(44):18060-18065.
63. Sayeed A, Konduri SD, Liu W, Bansal S, Li F, Das GM: **Estrogen receptor alpha inhibits p53-mediated transcriptional repression: implications for the regulation of apoptosis**. *Cancer Res* 2007, **67**(16):7746-7755.
64. Konduri SD, Medisetty R, Liu W, Kaipparettu BA, Srivastava P, Brauch H, Fritz P, Swetzig WM, Gardner AE, Khan SA *et al*: **Mechanisms of estrogen receptor antagonism toward p53 and its implications in breast cancer therapeutic response and stem cell regulation**. *Proc Natl Acad Sci U S A* 2010, **107**(34):15081-15086.
65. Jordan JJ, Inga A, Conway K, Edmiston S, Carey LA, Wu L, Resnick MA: **Altered-function p53 missense mutations identified in breast cancers can have subtle effects on transactivation**. *Molecular cancer research : MCR* 2010, **8**(5):701-716.
66. Rowan S, Ludwig RL, Haupt Y, Bates S, Lu X, Oren M, Vousden KH: **Specific loss of apoptotic but not cell-cycle arrest function in a human tumor derived p53 mutant**. *EMBO J* 1996, **15**(4):827-838.
67. Vaseva AV, Moll UM: **The mitochondrial p53 pathway**. *Biochim Biophys Acta* 2009, **1787**(5):414-420.
68. Heyne K, Schmitt K, Mueller D, Armbruester V, Mestres P, Roemer K: **Resistance of mitochondrial p53 to dominant inhibition**. *Mol Cancer* 2008, **7**:54.
69. Werner H, Karnieli E, Rauscher FJ, LeRoith D: **Wild-type and mutant p53 differentially regulate transcription of the insulin-like growth factor I receptor gene**. *Proc Natl Acad Sci U S A* 1996, **93**(16):8318-8323.
70. Gallagher EJ, LeRoith D: **Minireview: IGF, Insulin, and Cancer**. *Endocrinology* 2011, **152**(7):2546-2551.
71. Berns EM, Klijn JG, Look MP, Grebenchtchikov N, Vossen R, Peters H, Geurts-Moespot A, Portengen H, van Staveren IL, Meijer-van Gelder ME *et al*: **Combined vascular endothelial growth factor and TP53 status predicts poor response to tamoxifen therapy in estrogen receptor-positive advanced breast cancer**. *Clin Cancer Res* 2003, **9**(4):1253-1258.
72. Weisz L, Damalas A, Liontos M, Karakaidos P, Fontemaggi G, Maor-Aloni R, Kalis M, Levrero M, Strano S, Gorgoulis VG *et al*: **Mutant p53 enhances nuclear factor kappaB activation by tumor necrosis factor alpha in cancer cells**. *Cancer Res* 2007, **67**(6):2396-2401.
73. Roemer K: **Mutant p53: gain-of-function oncoproteins and wild-type p53 inactivators.** *Biol Chem* 1999, **380**:879-887.
74. Joshi H, Bhanot G, Borresen-Dale AL, Kristensen V: **Potential tumorigenic programs associated with TP53 mutation status reveal role of VEGF pathway**. *Br J Cancer* 2012, **107**(10):1722-1728.
75. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.
76. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer**. *Nat Genet* 2004, **36**(10):1090-1098.

77. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues**. *Bioinformatics* 2007, **23**(8):980-987.
78. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**(7074):353-357.
79. Bergmann S, Ihmels J, Barkai N: **Iterative signature algorithm for the analysis of large-scale gene expression data**. *Physical review E, Statistical, nonlinear, and soft matter physics* 2003, **67**(3 Pt 1):031902.
80. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS**. *BMC Bioinformatics* 2007, **8**(1):242.
81. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
82. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies**. *Proc Natl Acad Sci U S A* 2005, **102**(38):13544-13549.
83. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
84. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**(9):1943-1949.
85. Smyth GK: **Limma: linear models for microarray data**. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Duboit S, Irizarry R, Hubber W. New York: Springer; 2005.
86. Toft D, Shyamala G, Gorski J: **A Receptor Molecule for Estrogens - Studies Using a Cell-Free System**. *Proc Natl Acad Sci U S A* 1967, **57**(6):1740-&.
87. Kaufmann M, von Minckwitz G, Rody A: **Preoperative (neoadjuvant) systemic treatment of breast cancer**. *Breast* 2005, **14**(6):576-581.
88. Gennari A, Conte P, Rosso R, Orlandini C, Bruzzi P: **Survival of metastatic breast carcinoma patients over a 20-year period: a retrospective analysis based on individual patient data from six consecutive studies**. *Cancer* 2005, **104**(8):1742-1750.
89. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, Slamon DJ, Murphy M, Novotny WF, Burchmore M *et al*: **Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer**. *J Clin Oncol* 2002, **20**(3):719-726.
90. Arpino G, Gutierrez C, Weiss H, Rimawi M, Massarweh S, Bharwani L, De Placido S, Osborne CK, Schiff R: **Treatment of human epidermal growth factor receptor 2-overexpressing breast cancer xenografts with multiagent HER-targeted therapy**. *J Natl Cancer Inst* 2007, **99**(9):694-705.
91. Nahta R, Shabaya S, Ozbay T, Rowe DL: **Personalizing HER2-targeted therapy in metastatic breast cancer beyond HER2 status: what we have learned from clinical specimens**. *Current pharmacogenomics and personalized medicine* 2009, **7**(4):263-274.
92. Nahta R, Yuan LX, Zhang B, Kobayashi R, Esteva FJ: **Insulin-like growth factor-I receptor/human epidermal growth factor receptor 2 heterodimerization contributes to trastuzumab resistance of breast cancer cells**. *Cancer Res* 2005, **65**(23):11118-11128.

93. Munos B: **Lessons from 60 years of pharmaceutical innovation**. *Nature reviews Drug discovery* 2009, **8**(12):959-968.
94. Wiedswang G, Borgen E, Karesen R, Kvalheim G, Nesland JM, Qvist H, Schlichting E, Sauer T, Janbu J, Harbitz T *et al*: **Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer**. *J Clin Oncol* 2003, **21**(18):3469-3478.
95. Bergh J, Norberg T, Sjogren S, Lindgren A, Holmberg L: **Complete sequencing of the p53 gene provides prognostic information in breast cancer patients, particularly in relation to adjuvant systemic therapy and radiotherapy**. *Nat Med* 1995, **1**(10):1029-1034.
96. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET *et al*: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival**. *Proc Natl Acad Sci U S A* 2005, **102**(38):13550-13555.
97. 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**. *J Clin Oncol* 2009, **27**(8):1160-1167.
98. Kapp AV, Jeffrey SS, Langerod A, Borresen-Dale AL, Han W, Noh DY, Bukholm IR, Nicolau M, Brown PO, Tibshirani R: **Discovery and validation of breast cancer subtypes**. *BMC Genomics* 2006, **7**:231.
99. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V *et al*: **Inferring tumor progression from genomic heterogeneity**. *Genome Res* 2010, **20**(1):68-80.
100. Singh A, Settleman J: **EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer**. *Oncogene* 2010, **29**(34):4741-4751.
101. Li J, Wang K, Jensen TD, Li S, Bolund L, Wiuf C: **Tumor heterogeneity in neoplasms of breast, colon, and skin**. *BMC research notes* 2010, **3**:321.
102. Fan X, Shao L, Fang H, Tong W, Cheng Y: **Cross-platform comparison of microarray-based multiple-class prediction**. *PLoS One* 2011, **6**(1):e16067.
103. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome biology* 2010, **11**(8):R90.
104. Olivier M, Langerød A, Carrieri P, Bergh J, Klaar S, Eyfjord J, Theillet C, Rodriguez C, Lidereau R, Bièche I *et al*: **The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer.** *Clin Cancer Res* 2006, **12**:1157-1167.

I

Joshi et al. BMC Genomics 2012, 13:199
http://www.biomedcentral.com/1471-2164/13/199

BMC Genomics

RESEARCH ARTICLE Open Access

Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes

Himanshu Joshi[1,2], Silje H Nord[3], Arnoldo Frigessi[4], Anne-Lise Børresen-Dale[2,3] and Vessela N Kristensen[1,2,3*]

Abstract

Background: The human genome contains a large amount of *cis*-regulatory DNA elements responsible for directing both spatial and temporal gene-expression patterns. Previous studies have shown that based on their mRNA expression breast tumors could be divided into five subgroups (Luminal A, Luminal B, Basal, ErbB2$^+$ and Normal-like), each with a distinct *molecular portrait*. Whole genome gene expression analysis of independent sets of breast tumors reveals repeatedly the robustness of this classification. Furthermore, breast tumors carrying a *TP53* mutation show a distinct gene expression profile, which is in strong association to the distinct *molecular portraits*. The mRNA expression of 552 genes, which varied considerably among the different tumors, but little between two samples of the same tumor, has been shown to be sufficient to separate these tumor subgroups.

Results: We analyzed *in silico* the transcriptional regulation of genes defining the subgroups at 3 different levels: 1. We studied the pathways in which the genes distinguishing the subgroups of breast cancer may be jointly involved including upstream regulators (1st and 2nd level of regulation) as well as downstream targets of these genes. 2. Then we analyzed the promoter areas of these genes (−500 bp to +100 bp relative to the transcription start site) for canonical transcription binding sites using Genomatix. 3. We looked for the actual expression levels of the identified TF and how they correlate with the overrepresentation of their TF binding sites in the separate groups. We report that promoter composition of the genes that most strongly predict the patient subgroups is distinct. The class-predictive genes showed a clearly different degree of overrepresentation of transcription factor families in their promoter sequences.

Conclusion: The study suggests that transcription factors responsible for the observed expression pattern in breast cancers may lead us to important biological pathways.

Background

Previous studies have shown that breast tumors can be divided into five subgroups (Luminal A, Luminal B, Normal-like, ErbB2 over-expressing, and Basal-like) based on their mRNA expression patterns [1]. These patterns have been validated in independent datasets representing different laboratories, platforms and different patient cohorts [2]. Survival analyses on a sub-cohort of patients with locally advanced breast cancer showed a significant difference in outcome of the patients in the various expression subgroups, with poor prognosis for the ErbB2$^+$ and basal-like subtypes [2]. The expression of 552 genes, the *intrinsic gene list,* has been suggested to be sufficient to separate breast carcinomas into the five distinct subgroups. What mechanisms of common regulation make these genes cluster together? We have previously shown that we can separate the patient clusters based only on the promoter composition of single binding sites in the promoters of the genes from the intrinsic gene list [3]. However, regulation of gene expression in eukaryotes is highly complex and depends on sets of TFs rather than individual TFs [4] and in this study we attempt to characterize the overrepresentation of entire TF families. The promoter

* Correspondence: Vessela.Kristensen@medisin.uio.no
[1]Department of Clinical Molecular Biology and Laboratory Sciences (EpiGen), Division of Medicine, Akershus University Hospital, Lorenskog, Norway
[2]Institute for Clinical Medicine, University in Oslo, Oslo, Norway
Full list of author information is available at the end of the article

composition of the genes is one of the major determinants of gene regulation including multiple transcription binding sites that interact with a specific combination of transcription factors (TF). Eukaryotes achieve this diversity by combining a small number of transcription factors whose activities are modulated by diverse sets of conditions [5]. Different functionalities can be conferred on one TF by its association with different co-factors. These factors may act as global TFs that assist their gene-specific partners in their function, and may thus activate or repress transcription depending on the partner motif and the condition [5]. Analyzing transcription network dynamics in yeast, Luscombe et al. showed that, in response to diverse stimuli, transcription factors may alter their interaction patterns to varying degree, thereby rewiring the network [6]. While few transcription factors serve as permanent hubs, most of them act transiently during certain conditions. Exogenous processes like environmental responses facilitated fast signal transductions to multiple genes with short regulatory cascades, whereas endogenous processes needed to progress through multiple stages with a complex combination of TFs to fewer target genes [6]. The same TFs may act both in endogenous and exogenous processes. Regulatory hubs targeting disproportionately large numbers of genes and thereby representing the most influential components of a network- have been described. Both Pilpel [5] and Luscombe [6] concluded that precise regulation of a condition cannot arise from the specificity of individual TFs, therefore combinatorial TF usage seems to be the key. The NF-κB family of TFs is an example of transcription regulators that are activated by both intra- and extra-cellular stimuli such as cytokines, oxidant-free radicals, ultraviolet irradiation, and bacterial or viral products [7]. Aberrant NF-κB activity has been implicated in carcinogenesis and in the control of cellular response to anti-cancer agents. Activated NF-κB was detected predominantly in ER-negative breast tumors, and mostly in the ErbB2 over-expressing tumor subgroup [8].

Methods

The *in silico* analysis of the transcriptional regulation of genes defining the subgroups was performed at three different levels: (1) Study of the pathways in which the genes distinguishing the subgroups of breast cancer may be jointly involved including upstream regulators (1^{st} and 2^{nd} level of regulation) as well as downstream targets of these genes. (2) Then we analyzed the promoter areas of these genes (−500 bp to +100 bp relative to the transcription start site) for canonical transcription binding sites using Genomatix. (3) We looked for the actual expression levels of the identified TF and how they correlate with the overrepresentation of their TF binding sites in the separate groups.

Selection of genes

The expression of 552 genes, the *intrinsic gene list*, which has been suggested to be sufficient to separate breast carcinomas into the five distinct subgroups defined in [1] and [2,9] was used for the pathway analysis in this study (referred to as *full* list). A subset consisting of 197 genes [10] that best represented the classification scheme in breast cancer (referred as *top* list) were selected from the *intrinsic list*, and used in the promoter analysis part (Additional file 1: Table S1).

Pathway analysis

Pathway analysis was performed using Pathway Studio [11] from Ariadne Genetics. Two network prediction algorithms were used that allow to discover the patterns of gene expression inherent in the experimental data: Pearson Correlation and Auto Net Finder network prediction algorithm. Pathway Studio's text mining tools were applied to extract biological associations by mining PubMed to build pathways from extracted facts using data from recent publications and public and commercial databases such as KEGG, BIND, GO, and the PathArt database of curated signaling and disease pathways. The algorithm for building Correlation Network in Pathway Studio is based on Pearson Correlation. Genes with similar expression profiles are connected with edges indicating the significance of the correlation. The group of tightly correlated genes form cluster in the correlation network. The algorithm can be used for clustering genes according to their expression profiles across multiple samples. The tool calculates correlation coefficients between all pairs of gene expression profiles measured in the experiment and outputs clusters of highly correlated genes. Identified gene clusters can be further validated and analyzed using relations from the database that have been extracted from the literature by Ariadne Genetics. Auto Net Finder is a network estimation system that combines hierarchical clustering and Graphical Gaussian Modeling and is used for distinguishing direct and indirect relationship among variables. Bibliosphere pathways (release 7.1) [12] (*http://www.genomatix.de*, Genomatix Software GmbH) was used for extracting the associations between gene, transcription factor and proteins corresponding with the genesets defining each molecular subtype of breast cancer. Genomatix Bibliosphere is a knowledge database consisting of manually curated co-cited genes in PubMed, which additionally provides information about the presence of TFBS in their promoters, using *in silico* tool- MatInspector, interactions and associated pathways from Molecular Interactions database-NetPro and BioCyc, respectively.

Analysis of overrepresentation of TFBS families in the promoter sequences

We extracted the putative regulatory promoter regions from 500 bp upstream to 100 bp downstream of RefSeq promoters of the subtype-associated genes. Further analysis was based on the hypothesis that overrepresentation of potential transcription factor binding site (TFBS) motifs in a set of co-expressed gene promoters may indicate regulatory relationship. In order to emphasize the functional representation of TFBS motifs overrepresented in a set of promoters, we used the TFBS matrix family concept. TFBS matrix families are defined as groups of TFBS weight matrices corresponding to the same or functionally similar transcription factors. For any given TF, there could be multiple matrices described by different independent sources, leading to multiple matches for similar position or shifting of matches by a few base pairs. By using the functional domain clustering based on di/tri/tetra-nucleotide occurrence and additionally function-based subgrouping, TFBS matrices can be grouped according to their functional similarity, known as TFBS families [13]. Thus members sharing same TFBS family are expected to have functional similarity in addition to binding domain similarity. For estimation of over-representation of each TFBS family, first occurrences of its corresponding TFBS motifs within a set of subtype-specific promoter sequences was obtained. Then relative occurrence of each TFBS family was estimated by comparing this observed occurrence to the rate of occurrence of the same TFBS matrix family in an equal base-pair long reference background sequences from human promoter. Overrepresentations of a motif is measured by two different methods:

1. In terms of fold factor of overrepresentation compared to the background
   Fold factor of TFBS overrepresentation was calculated by a formula as mentioned below:

$$r(X) = \frac{n_{obs}(X)}{n_{exp}(X)}$$

   Where, $r(X)$ = fold factor of overrepresentation of a TFBS family, X
   n_{obs} (X) = observed number of hits of X in a given set of promoter sequences
   n_{exp} (X) = expected number of hits of X in an equally sized sample from genomic promoter background sequences
2. As z-scores that provide a measure of the distance of sample from the reference population mean. Here sample refers to the number of observed hits of any particular TFBS in a given input set of sequences and reference refers to the number of hits of the same TFBS in equally sized human genomic promoter sequence population.

$$z(X) = \frac{n_{obs}(X) - n_{exp}(X) - 0.5}{S(X)}$$

   $z(X)$ is a z-score of overrepresentation of a transcription factor binding site family (X);
   n_{obs} (X) is a number of observed hits of X in an input promoter sequences;
   n_{exp} (X) is expected number of hits of X in an equally sized sample sequences in human genomic promoter background;
   $S(X)$ is a population standard deviation of number of hits of X

We used Genomatix RegionMiner tool (Genomatix Software GmbH, *http://www.genomatix.de*) in order to evaluate the degree of TFBS family overrepresentation. The histogram of z-scores of each TFBS motif families in each subtype-specific promoter sequences is shown in the Additional file 2: Figure S1. Histograms like this indicate that choosing the cut-off level of 2.0 allows identifying TFBS families that are overrepresented. However, z-score cut-off level of 2.0 does not provide a precise measure of significance, because of the disparity of sample size between sample and reference. Due to the copyright and technical limitations in accessing the Transfac database, further statistical testing of over-representation could not be performed within that tool.

Under-representations or absence of TFBS family motifs in sub-type specific genes may occur due to a fewer number of subtype-representative genes and subsequently a smaller number of promoter sequences used for any particular subtype. This can be a source of false positivity. Therefore we have not taken into account the under-representations of TFBS family motifs in this analysis.

Principal component analysis to identify TFBS with maximum variance between subtypes

Principal component analysis (PCA) [14] was performed for ranking the TFBS families with respect to the variance of fold-factor overrepresentation contributed by them between five subtypes. We prepared a matrix of TFBS fold-factors for subtypes, with subtypes as columns and TFBS families as rows. We performed PCA on this matrix using the *princomp* function of *Matlab*. Subtracting each data point from the column mean represents a center of this matrix. Hotelling's T^2 statistic was used as a

measure of multivariate distance of each TFBS family from the center of the TFBS fold-factor matrix as described in [*http://www.mathworks.com/help/toolbox/stats/princomp.html*].

Gene expression data

We used a subset of the samples ($n = 114$) from previously published [15] mRNA expression data [GEO dataset #GSE19783]. Subtypes were predicted by using the *PAM50* [16].

mRNA expression of the studied TF

Transcription factor families with overrepresentation *z*-score >2.0 were mapped to their corresponding probes in the mRNA expressions dataset. By applying multiclass SAM, we extracted 120 TF genes with significantly different (at the FDR <0.1) expression between the five subtypes. Pearson's correlation between the subtype-specific geometric mean expression of this subset of transcription factor genes and fold overrepresentation was computed. The justification of using geometric mean instead of arithmetic mean is that typically mRNA expression values are log-normally distributed.

Results and discussion

Pathway analysis of the genes that define the five breast cancer subgroups

Using Pathway Studio from Ariadne Genetics, we studied the direct interactions between the genes with distinguished gene expression pattern in the breast cancer subgroups as described in *Materials and Methods, selection of genes*. Most profound direct interactions were observed for the genes defining the luminal A group with protein-protein interactions between *XBP1* and *ESR1* and *CCND1* (Additional file 3: Figure S2). Trefoil *(TFF3*) has been functionally coupled to *CCND1* through angiotensin receptor 1 (*AGTR1*). Angiotensin II is converted from its precursor by angiotensin I-converting enzyme (ACE) and has been shown to mediate growth in breast cancer cell lines via ligand-induced activity through the angiotensin II type 1 receptor (*AGTR1*). We also searched for upstream regulators as well as downstream targets of these genes. Downstream targets could be observed centered at the ESR1, MYC, NFKB1, GATA3, CCND1, TP53 and MSX2/FOXC1 (Additional file 4: Figure S3).

A somewhat less organized pathway structure is observed in the luminal B subclass. The ESR1 node was not observable and the TP53 network was more sparse with fewer partner genes. Novel nodes were centered at NRG1, GSTP1 and CUL1 (Additional file 5: Figure S4), CUL1 has homology to yeast Cdc53, which is part of a complex known as SCF that mediates the ubiquitin-dependent degradation of G1 cycles and cyclin-dependent kinase inhibitors, while NRG1 contains a domain related to the epidermal growth factor family of ligands and can act as receptor agonists. The direct interactions between genes highly expressed in Luminal B subtype were observed between *GSTP1* and *CDK2AP1*, *S100A10* and *S100A11* and *PPP1R13B* and *TP53BP2*. The latter protein interacts with *TP53* to specifically enhance p53-induced apoptosis but not cell cycle arrest.

Four distinct regulatory nodes were observed in the ERBB2 group: around the ERBB2 itself, TP53, NFKB1 and CTNNB1 (cadherin-associated protein, beta 1) (Additional file 6: Figure S5). NFkB-p65 was shown to repress β-catenin-activated transcription of cyclin D1 [17]. Moreover, a direct interaction is established between *ERBB2* and *GRB7* (Additional file 3: Figure S2). The solution structure of the Grb7-SH2/erbB2 peptide complex was described and suggested to be involved in cell signaling pathways that promote the formation of metastases and inflammatory responses. *PPARBP*, which is co-amplified with ERBB2, has in early studies been suggested to play a role in mammary epithelial differentiation and in breast carcinogenesis by its ability to function as *ESR1* coactivator. It was shown to contain a typical CCAT box and multiple cis-elements such as C/EBPbeta, YY1, c-ETS-1, AP1, AP2, and NFkappaB binding sites. The 4 different regulatory nodes are connected by *FLOT2*, the human epidermal surface antigen involved in epidermal cell adhesion. *NFKB1* was present in the network for the Basal group, where also the FOX family, a whole family of cyclins and CDK2, and CDK6 and isoforms of protein kinase (*RPS6K*) were present (Additional file 7: Figure S6). Interestingly, a large number of connections lead to *GJA1* (Cap junction protein, alpha, also known as connexin 43). Other distinct nodes around TP53 are those connecting to KRT5, MAPK signalling, E2F1 and NCL. NCL, Nucleolin, one of the most abundant nucleolar proteins, has been recently shown to be involved in the reprogramming of somatic cells for derivation of either embryonic stem (ES) cells, by somatic cell nuclear transfer (SCNT), or ES-like cells, by induced pluripotent stem (iPS) cell procedure. Nucleolar proteins are proposed to be the markers of activation of embryonic genes [18] and provide mechanism for nucleolar control of progression of cell cycle in stem cells and cancer cells [19]. TP53 was a central node in the regulatory network of the normal-like subgroup, surrounded by JUN, ACSS2, ACSL1, KRT13, PIK3R1 and other nodes some representing glycolysis, energy metabolism, pyruvate metabolism and metabolism of *carbohydrate* (Additional file 8: Figure S7).

Noteworthy, a TP53 network node was observed in each of the studied expression subclasses shown here (Additional file 4: Figure S3, Additional file 8: Figures S7). It is of interest to note that in every case TP53 was a hub in a somewhat different neighborhood. While in

the basal subtype TP53 was connected to CDK6, a cyclin-dependent protein kinase (CDKs) that regulate major cell cycle transitions and CDH3, cadherin 3, as well as FZD7 and KRT5, in the luminal A tumors one could observe detoxifying enzymes such as NAT1, CYP2A6 as well as the retinoic acid receptor RARRES3 in the TP53 hub (Figure 1).

Over-representation of specific transcription factor binding sites in the promoter of the genes that distinguish the subtypes

The correlation matrix of TFBS fold-overrepresentation vectors for the five subtypes shows positive correlation in terms of potential TFBS family overrepresentation between 1. ERBB2+ and basal subtypes (0.27); 2. Luminal B and ERBB2+ (0.16); 3. Luminal A and luminal B (0.11). In order to visualize the differential TFBS overrepresentation, we performed the principal component analysis (PCA). PCA plot (Figure 2) displays the significant differences between the subtypes in terms of fold-factor of motif frequencies observed in promoter sequences of subtype-associated gene promoters compared to their corresponding normal frequencies in genomic promoter sequences. Distances between points representing the TFBS matrix families are the multivariate distances of fold-factor overrepresentation of each TFBS family in each of the subtype. This indicates that the shorter the distance, the greater similarity in fold-overrepresentation of that particular TFBS family in given subtypes. More than 60% and 76% of cumulative variance is captured by first two components and first three principal components, respectively. The top ten ranking TFBS families in distance from center and some of the functionally significant TFBS families are specifically labeled in the PCA plot. Biplots of first and second principal components show differentially overrepresentated TFBS families between the normal-like and rest of the subtypes. Biplot of second and third principal components shows TFBS family overrepresentations in luminal B. Differential TFBS family representations between ERBB2+ and basal groups cannot be seen in biplots of first three principal components, but can be visualized in a biplot of first and fourth principal components. In the first principal component, V$BTBF, V$PAX1, V$PAX4 and V$TCFF

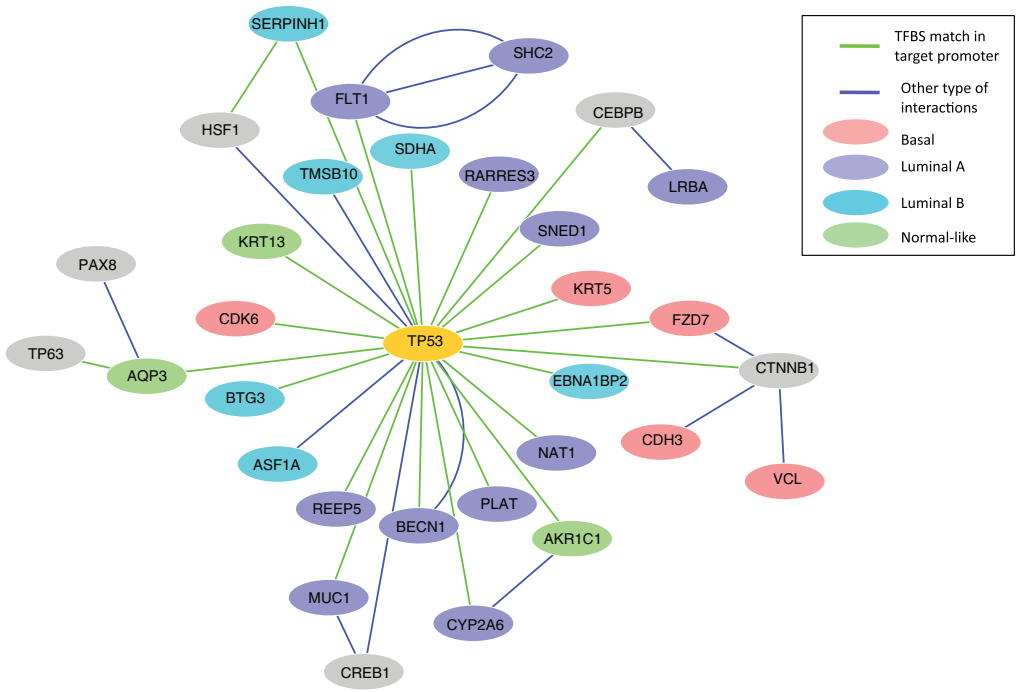


Figure 1 Predicted functional relationship of TP53 in different molecular subtypes of breast cancer. Figure shows predicted interactions of genes or proteins with TP53. Source: *Bibliosphere pathway database*. (*green edges*: TF motif match found in target promoter of target genes; genes associated with basal subtype are shown as *red* nodes, ones with luminal A in *blue*, luminal B in *cyan* and normal-like as *green* nodes.)

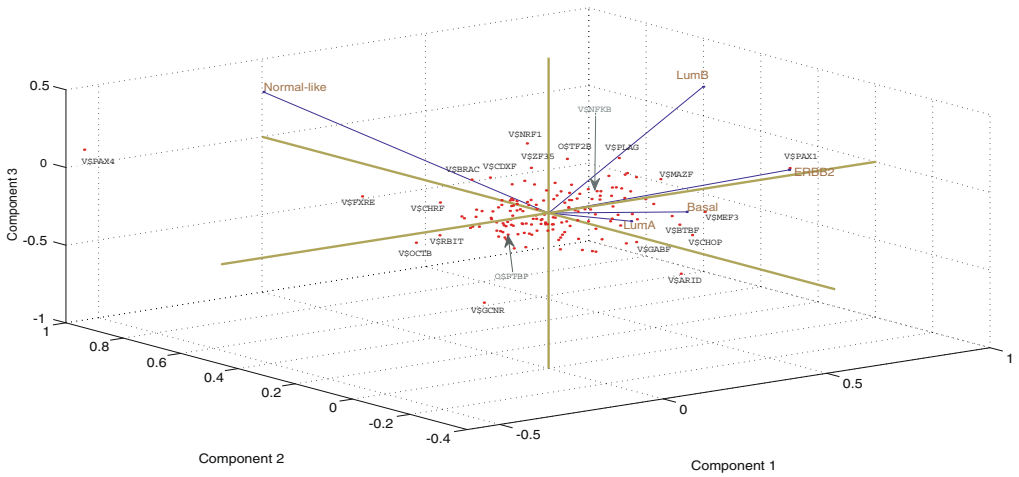


Figure 2 PCA plot of overrepresentated TFBS matrix families. PCA plot shown in terms of fold factor of overrepresentation in each subtype compared to the reference genomic promoters background. Blue lines represent the Eigen vectors, the direction and length of which indicates how each subtype variable contributes to the principal components in the plot. TFBS matrix families with maximum distance from the centroid are labeled on the plot.

are the major contributors of variance, where as V$PAX4, V$GUCE, V$ARID are the major contributors of variance in the second principal component.

Several of the gene clusters shared *cis*-elements that were present in more than 90% of the promoters. For the top six genes that classify the ErbB2+ over-expressing cluster, four TFBSs were found to be present in 100% of the promoters. These were NOLF (Neuron-specific-olfactory), ETSF (E26 Transformation-Specific factor 1), STAT (the Signal Transducers and Activator of Transcription protein) and NF-κB (Nuclear Factor κappa Beta) (Additional file 9: Table S2). NF-κB is the family of nuclear factor kappa beta of transcription factors. NF-κB has been shown to promote cell proliferation, to suppress apoptosis, to promote cell migration, and suppress differentiation [7]. NF-κB binding sites were found significantly over-represented in the promoters that best classify the ErbB2$^+$ subgroup compared to the other 4 subgroups (Additional file 9: Table S2; Figure 3B) and 78% of the 27 genes expressed in the basal-like subgroup had also NF-κB binding site in the promoter. This was in marked contrast compared to the promoter composition of the normal-like and luminal subgroups (Figure 3B). The presence of NF-κB binding sites in the genes from the ERBB2 and basal groups is in concordance with the pathway analysis performed on the downstream genes (see above). The *cis*-elements PAX1, PAX9 (The paired box gene 5), MAZF (*myc*-associated zinc finger) and EGRF (epidermal growth factor receptor) were overrepresented in the genes that are over-

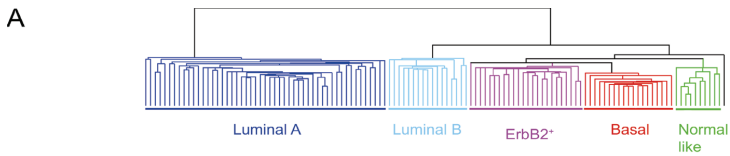


Figure 3 Subtypes with relevance to NF-κB binding sites and TP53 mutations. A. The five subtypes shown by hierarchical clustering using the "intrinsic" gene set. Dendrogram shows the clustering of the tumors into five subgroups. Branches are color-coded. **B.** Frequency of NF-κB binding sites in the 5 subgroups; and **C.** Frequency of *TP53* mutations.

expressed in the Luminal B subgroup (Additional file 9: Table S2). While the PAX superfamily is involved in a multitude of developmental processes and is required for initiating B cell lineage and maintaining neural development and spermatogenesis, the MAZF is a common transcription factor and might play a more general role. The major distinction between the luminal A and B, both consisting of ER positive tumors, is the presence of a strong proliferations cluster in the luminal B subtype. Noteworthy, binding sites for growth factors and their receptors like EGRF are over-represented in the promoters of the genes that define the luminal B subgroup and were overrepresented in the pathway analysis as well *(see above)*. EGRF is not only a receptor for EGF (Epidermal growth-factor), but also for other members of the EGF family and it is involved in the control of cell growth and differentiation. For the geneset of the normal-like subgroup, we observed overrepresentation of NRF1 family of TFBS (Additional file 9: Table S2).

Presence of promoter modules in genes that define the ErbB2+ subgroup

The specificity of promoter-controlled gene regulation may depend on the relative organization of the elements within the promoter rather than solely on individual elements [20–22]. Genes expressed in the same functional context do often share promoter modules [20,21]. The binding elements are often occupied differently in different tissues, and these differences can be used to derive all type-specific sub-modules *in silico*. A promoter module may be defined as an organized group of regulatory elements where both order and distance should be considered. Genes expressed in the same functional context do often share promoter modules [20,21]. For the six best genes of the ErbB2+ over-expressing cluster, a common framework consisting of NF-κB and ETS1 transcription factor binding sites was found (Figure 4). The ETS are fundamentally important TFs with roles in cell development, cell differentiation, cell proliferation, apoptosis and tissue remodeling (reviewed [23]). The family is characterized by an evolutionarily conserved DNA-binding domain that regulates expression by binding to a purine-rich core sequence in cooperation with other TFs. Most of the proteins in the ETS family are downstream nuclear targets of *ras*-MAP kinase signaling, and the deregulation of ETS genes results in the malignant transformation of cells [24] It has previously been reported that mutant TP53 required ETS1 to synergistically activate the expression of *ABCB1*. ETS1 was shown to interact exclusively with mutant TP53 in vivo, but not with wild-type TP53 [25]. High levels of ETS1 expression were associated with poorer prognosis [26]. The presence of a promoter module constituting of NF-κB and ETS has been reported previously in genes co-regulated

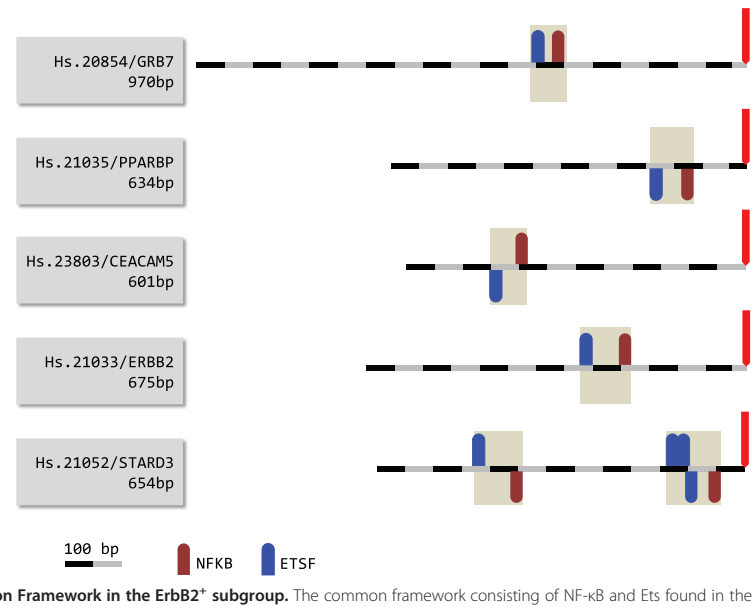


Figure 4 Common Framework in the ErbB2^{+} subgroup. The common framework consisting of NF-κB and Ets found in the 6 cluster defining genes of the ErbB2+ over-expressing subgroup. Distance to next element is between 29 and 79 bp (ETSF). Directions (up/down) of the elements indicate presence of hits on sense or antisense strands respectively.

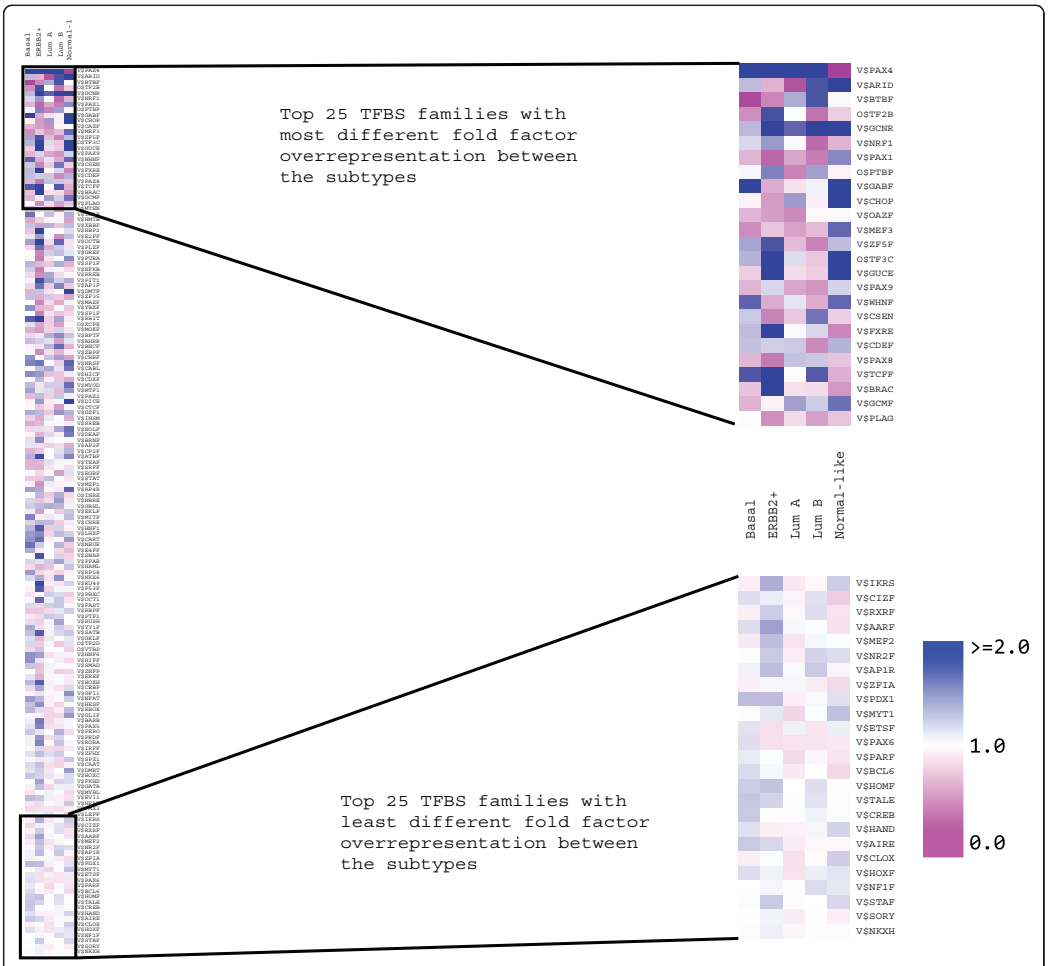


Figure 5 Heatmap of fold overrepresentations for the TFBS families ranked according to their distance from center. Out of all, 25 TFBS matrix families top ranked according to their distance from the centroid of fold-overrepresentation matrix, are highlighted.

in mitogen-stimulated T-cells [27]. Interactions between members of the ETS family and NF-κB have been described previously. ETS1 induces IKKα expression. IKKα is a kinase that marks the NF-κB inhibitor IκB for degradation, and active NF-κB is translocated to the nucleus. ETS1-mediated activation of IKKα is negatively regulated by TP53 binding to ETS1. TP53 physically interacts with ETS1 and specifically inhibits ETS1 induced IKKα promoter activity. Loss of TP53-mediated control over ETS1 dependent transactivation of IKKα may represent a novel pathway for the constitutive activation of NF-κB mediated gene expression and therapy resistance in cancer cells [28] TP53 is therefore an ETS1 and ETS2 target gene [29]. NF-κB controls a broad spectrum of genes by a variety of mechanisms in response to diverse environmental changes. NF-κB may be a universal regulator, while ETS could reflect cell-type or stimulation specific differences since ETS binding sites were detected in a fraction of the NF-κB controlled genes.

Over-representation of TP53 mutations in the tumors that belong to the ErbB2$^+$ and basal-like subgroups

In human breast tumors, the two tumor subgroups exhibiting the most prominent activation of putative NF-κB target genes (ErbB2$^+$ and Basal-like) also harbored the

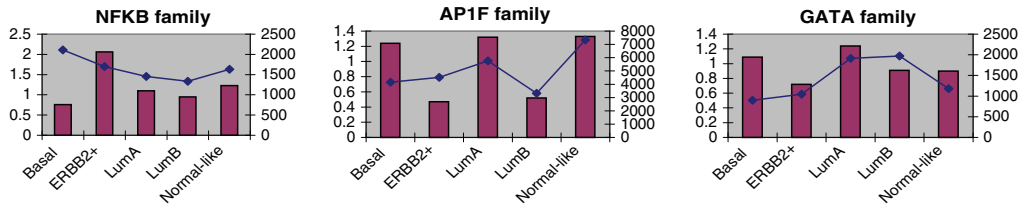


Figure 6 Correlation between TF overrepresentation and corresponding TF gene in each subtype. Correlation between geometric mean of TF gene expressions in each subtype (shown as red bars) and their corresponding TF matrix family overrepresented in subtype-specific promoter sequences (shown as blue points) is plotted.

highest frequency of p53 mutations. 86% of the patients in the ErbB2$^+$ subgroup had *TP53* mutations in their tumors and all the genes that are abnormally expressed in this tumor type have NF-κB binding sites in their promoter (Figure 3C). There is an evidence that NF-κB can regulate *TP53* expression and that NF-κB is required for TP53-dependent cell death [30]. In turn, TP53 activates NF-κB through the RAF/MEK1/p90 pathway [30]. The TP53 protein interacts with NF-κB and enhances its transcriptional activity and its anti-apoptotic efficacy. Over-expression of ErbB2 is known to induce the classical NF-κB pathway [31,32]. The estrogen receptor (ER) can bind physically to NF-κB to inhibit its DNA binding functions, hitherto repressing gene expression [33]. Therefore the NF-κB pathway was shown to be a major stroma-tumor signaling mediator in ER negative tumors with over-expression of ErbB2 [8]. NF-κB signaling has been associated with doxorubicin resistance, and agents blocking NF-κB function have been proven beneficial in the treatment of tumors in combination with standard anti-cancer therapies [34].

Over-represented transcription factor families within the promoter sequences

We observed the over-representation of V$BTBF (*kaiso*), V$OAZF and V$PAX8 in basal and ERBB2+ tumor associated gene promoters (Figure 5, Additional file 10: Table S3). *Kaiso* group of transcription factors are known to show nuclear accumulation during active mitosis [35] and their over-representation indicates potential functional role in these two subtypes showing aggressive tumor progression and high cell proliferation. PAX8 activity has also been observed in metastatic renal tumors [36]. Precise role of PAX8 and OAZF groups of transcription factors is yet unknown in breast cancers. ERBB2+ gene promoters also show over-representation of V$NFKB, Pleomorphic adenoma gene associated V$PLAG and *ras*-responsive element binding protein associated V$RREB families of TFBS. Activity of NFKappa B is already discussed in the earlier section. RREB1 activity plays a role in TP53 mediated apoptosis [37] that gets perturbed in absence of functional TP53, which is a common phenomenon in ERBB2+ tumors. Both luminal groups involve over-representation of PAX subgroup 1 member TFBS's- V$PAX1, V$PAX9 and V$ZF5F families. PAX9 activity is known to be a marker of better prognosis. Overrepresentation of V$P53F, V$HOXF, V$CLOX, V$PARF and V$GATA was observed specifically in luminal A group in which estrogen receptor signaling is a predominant characteristic. The transcription factors corresponding to V$PARF group (PAR bZIP TFs) are mediators in oxidative stress-induced apoptosis [38]. In the luminal B group of promoters, we observed over-representation of V$EGRF, V$CTCF and V$EKLF etc. Egr-1 which corresponds to the V$EGRF family is known to be associated with cell cycle entry in response to growth stimuli [39]. We also observed significant over-representation of V$NRF1 in both normal-like and luminal B group of promoters. NRF-1 transcription factor is an oxidant-sensitive transcription factor, usually found in ER positive breast cancers [40] and is shown to be associated with higher tumor grade [41].

By using the Wilcoxon rank sum test, we observed significantly elevated mRNA expressions of *ESR1* and *PGR* in Luminal A or Luminal samples compared to the basal ones ($p < 1.0e-6$), with non-significant differences in *ERBB2* expressions. As expected *ERBB2* was significantly upregulated in ERBB2+ tumors along with downregulated *ESR1* and *PGR*, compared to the rest ($p < 1.0e-4$). Regulation by many transcription factors shown overrepresented here in ER + ve or ER-ve subtypes is not well characterized in context of estrogen and progesterone receptor activity. However, overrepresentation of some of the TFBS, such as GATA, BTBF, NF Kappa B – appear to be consistent with prevailing knowledge about the subtypes and their ER/PR or Her2 status.

Thus functions of the TF genes corresponding to the over-represented TFBS families hint the predominant characteristics of the subtypes. Findings from the above *in silico* analysis will be further validated in reporter studies and ChIP analyses. The approach of identifying

overrepresented TFBS in a set of coordinately expressed genes under a particular disease class or condition can improve the specificity and noise tolerance [42]. However, its main limitation is that it does not account for the role of local chromatin environment constituted by structural properties, epigenetic modification etc. The local chromatin environment can offer condition-specific functionality to the existing TFBSs in a set of promoters.

Promoter sequences extending from 500 bp upstream to 100 bp downstream relative to TSS typically contain core promoter elements, CpG islands, downstream promoter element and other components of transcriptional machinery. Besides, this region has been demonstrated to have high density of positional as well as comparative TFBS [43], many of which are typically location sensitive. Thus limiting the analysis to this proximal promoter region, rather than analyzing the broader region (i.e. -1000 bp to +500 bp relative to the TSS) – could reduce false positives in TFBS overrepresentation. However, by that very limitation we may omit important information about second alternative promoters and distant control loci, which are therefore outside the scope of this analysis.

Correlation between actual abundance of TFs and frequency of their BS in the genes defining the clusters

Some of the TFBS family overrepresentations were positively correlated with the geometric means of subtype-specific mRNA expressions of their corresponding TF genes. (Shown in Figure 6, Additional file 11: Table S4). The rationale underlying the use of geometric mean is that gene expression intensity values follow lognormal distribution.

Biological uncertainty in a correlation between the abundance of TFs and frequency of their BS might be attributed to several factors. The most common and obvious reason could be mutant or copy number altered TF. Moreover, here we have not accounted for the expressions of downstream targets of the TFs. It is noteworthy that mutations (point mutation and copy number alteration) in TFs can also have an impact on the level of expression of the downstream genes. For instance, a mutant TP53, which is still highly expressed, may not recognize the original binding sites anymore, leading to a drop in the expression of the target genes.

Conclusion

Here we report that the promoter composition of the genes that strongly predict the patient subgroups is distinct. The gene classes showed a clear separation when based solely on their promoter composition. This finding suggests that studying those transcription factors associated to the observed expression pattern in breast cancers may lead us to important biological pathways responsible for the regulation of gene expression in breast cancer.

Additional files

Additional file 1: Table S1. Subtype-specific gene list. Table shows the 197 subtype-specific best discriminatory genes, which is a subset of the intrinsic gene-list.

Additional file 2: Figure S1. Histogram of z-scores of overrepresentation. Histogram of TFBS matrix family overrepresentation observed in subtype-specific promoters compared to the reference genomic promoter background shown as z-scores.

Additional file 3: Figure S2. Direct interactions between genes defining subtypes. Subtype-relevant key driver interactions for Luminal A, B and ERBB2+ subtypes.

Additional file 4: Figure S3. Protein-protein interactions and TF interactions associated with Luminal A subtype. Network shown here is based on the luminal A specific genelist.

Additional file 5: Figure S4. Protein-protein interactions and TF interactions associated with Luminal B subtype. Network shown here is based on the luminal B specific genelist.

Additional file 6: Figure S5. Protein-protein interactions and TF interactions associated with ERBB2+ subtype. Network shown here is based on the ERBB2+ subtype-specific genelist.

Additional file 7: Figure S6. Protein-protein interactions and TF interactions associated with basal subtype. Network shown here is based on the basal subtype-specific genelist.

Additional file 8: Figure S7. Protein-protein interactions and TF interactions associated with normal-like subtype. Network shown here is based on the normal-like subtype-specific genelist.

Additional file 9: Table S2. TFBS overrepresentation in subtypes-specific gene promoters. List of significantly over-represented transcription factor binding site families in subtypes of breast cancers at the cut-off level of z-score > =2.0.

Additional file 10: Table S3. Over-representation of potential TFBS in subtype-specific promoter sequences. Table shows the fold over-representation of potential transcriptional factor hits (represented as TFBS families) in subtype- specific gene promoter sequences.

Additional file 11: Table S4. Correlation between TFBS overrepresentation and mRNA expression of corresponding TF genes. Table displays the Pearson's correlation between the geometric mean of expression values of transcription factor genes in subtypes and fold overrepresentation of corresponding TFBS families.

Abbreviations

TF: transcription factor; TFBS: transcription factor binding site; PCA: principal component analysis; ER: estrogen receptor; PGR: progesterone receptor; Her2: Human Epidermal Growth Factor Receptor 2.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VNK conceived and designed the study and helped to draft the manuscript. AF provided statistical expertise into the methods used in this study. HJ performed TFBS overrepresentation analysis, wrote the corresponding sections of manuscript, prepared figures and tables and revised the manuscript. SHN participated in pathway analysis using Pathway Studio tool and performed promoter module analysis using MatInspector. VNK and AF approved the final manuscript. All authors read and approved the final manuscript.

Authors' information

HJ is a fellow of the Health Authority of South-East Norway. SHN is a fellow of the Norwegian cancer society (Den Norske Kreftforening).

Acknowledgements

This work was supported by grants D-99061 and D-03067 from The Norwegian Cancer Society, grant 152004/150 from The Functional Genomics program (FUGE), The Norwegian Research Council (NFR) and grant 155218/300 from NFR.

Author details
[1]Department of Clinical Molecular Biology and Laboratory Sciences (EpiGen), Division of Medicine, Akershus University Hospital, Lorenskog, Norway. [2]Institute for Clinical Medicine, University in Oslo, Oslo, Norway. [3]Department of Genetics, Institute for Cancer Research Oslo University Hospital, Radiumhospitalet, Norway. [4]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Blindern, 0317, Oslo, Norway.

Received: 14 October 2011 Accepted: 17 February 2012
Published: 22 May 2012

References

1. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747–752.
2. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale A-L, Botstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100:**8418–8423.
3. Tongbai R, Idelman G, Nordgard SH, Cui W, Jacobs JL, Haggerty CM, Chanock SJ, Børresen-Dale A-L, Livingston G, Shaunessy P, Chiang C-H, Kristensen VN, Bilke S, Gardner K: **Transcriptional networks inferred from molecular signatures of breast cancer.** *Am J Pathol* 2008, **172:**495–509.
4. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.** *Genome Res* 2003, **13:**773–780.
5. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153–159.
6. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431:**308–312.
7. Chen F, Castranova V, Shi X: *New insights into the role of nuclear factor-kappaB in cell growth regulation.* 2001, **159:**387–397.
8. Biswas DK, Shi Q, Baily S, Strickland I, Ghosh S, Pardee AB, Iglehart JD: **NF-kappa B activation in human breast cancer specimens and its role in cell proliferation and apoptosis.** *Proc Natl Acad Sci USA* 2004, **101:**10137–10142.
9. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Van De Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale A-L: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869–10874.
10. Muggerud AA, Johnsen H, Barnes DA, Steel A, Lønning PE, Naume B, Sørlie T, Børresen-Dale A-L: **Evaluation of MetriGenix custom 4D™ arrays applied for detection of breast cancer subtypes.** *BMC Cancer* 2006, **6:**59.
11. Nikitin A: **Pathway studio–the analysis and navigation of molecular networks.** *Bioinformatics* 2003, **19:**2155–2157.
12. Scherf M, Epple A, Werner T: **The next generation of literature analysis: Integration of genomic analysis into text mining.** *Brief Bioinform* 2005, **6:**287–297.
13. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21:**2933–2942.
14. Jolliffe IT: **Principal Component Analysis.** *Chemom Intell Lab Syst* 1986, **2:**37–52.
15. Enerly E, Steinfeld I, Kleivi K, Leivonen S-K, Aure MR, Russnes HG, Rønneberg JA, Johnsen H, Navon R, Rødland E, Mäkelä R, Naume B, Perälä M, Kallioniemi O, Kristensen VN, Yakhini Z, Børresen-Dale A-L: **miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors.** *PLoS One* 2011, **6:**13.
16. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160–1167.
17. Hwang I, Choi YS, Jeon M-Y, Jeong S: **NF-κB p65 represses β-catenin-activated transcription of cyclin D1.** *Biochem Biophys Res Commun* 2010, **403:**79–84.
18. Johansson H, Svensson F, Runnberg R, Simonsson T, Simonsson S: **Phosphorylated nucleolin interacts with translationally controlled tumor protein during mitosis and with Oct4 during interphase in ES cells.** *PLoS One* 2010, **5:**e13678.
19. Tsai RYL, McKay RDG: **A nucleolar mechanism controlling cell proliferation in stem cells and cancer cells.** *Genes Dev* 2002, **16:**2991–3003.
20. Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel AE: **COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation.** *Nucleic Acids Res* 2000, **28:**311–315.
21. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15:**180–186.
22. Fessele S, Maier H, Zischek C, Nelson PJ, Werner T: **Regulatory context is a crucial part of gene function.** *Trends Genet* 2002, **18:**60–63.
23. Oikawa T, Yamada T: **Molecular biology of the Ets family of transcription factors.** *Gene* 2003, **303:**11–34.
24. Oikawa T: **ETS transcription factors: possible targets for cancer therapy.** *Cancer Sci* 2004, **95:**626–633.
25. Sampath J, Sun D, Kidd VJ, Grenet J, Gandhi A, Shapiro LH, Wang Q, Zambetti GP, Schuetz JD: **Mutant p53 cooperates with ETS and selectively up-regulates human MDR1 not MRP1.** *J Biol Chem* 2001, **276:**39359–39367.
26. Dittmer J: **The Biology of the Ets1 Proto-Oncogene.** *Mol Cancer* 2003, **2:**29.
27. De Siervi A, De Luca P, Moiola C, Gueron G, Tongbai R, Chandramouli GVR, Haggerty C, Dzekunova I, Petersen D, Kawasaki E, Kil WJ, Camphausen K, Longo D, Gardner K: **Identification of new Rel/NFkappaB regulatory networks by focused genome location analysis.** *Cell cycle Georgetown Tex* 2009, **8:**2093–2100.
28. Gu L, Zhu N, Findley HW, Woods WG, Zhou M: **Identification and characterization of the IKKalpha promoter: positive and negative regulation by ETS-1 and p53, respectively.** *J Biol Chem* 2004, **279:**52141–52149.
29. Sementchenko VI, Watson DK: **Ets target genes: past, present and future.** *Oncogene* 2000, **19:**6533–6548.
30. Ryan KM, Ernst MK, Rice NR, Vousden KH: **Role of NF-kappaB in p53-mediated programmed cell death.** *Nature* 2000, **404:**892–897.
31. Guo G, Wang T, Gao Q, Tamae D, Wong P, Chen T, Chen W-C, Shively JE, Wong JYC, Li JJ: **Expression of ErbB2 enhances radiation-induced NF-kappaB activation.** *Oncogene* 2004, **23:**535–545.
32. Pianetti S, Arsura M, Romieu-Mourez R, Coffey RJ, Sonenshein GE: **Her-2/neu overexpression induces NF-kappaB via a PI3-kinase/Akt pathway involving calpain-mediated degradation of IkappaB-alpha that can be inhibited by the tumor suppressor PTEN.** *Oncogene* 2001, **20:**1287–1299.
33. Ray P, Ghosh SK, Zhang DH, Ray A: **Repression of interleukin-6 gene expression by 17 beta-estradiol: inhibition of the DNA-binding activity of the transcription factors NF-IL6 and NF-kappa B by the estrogen receptor.** *FEBS Lett* 1997, **409:**79–85.
34. Wang CY, Cusack JC, Liu R, Baldwin AS: **Control of inducible chemoresistance: enhanced anti-tumor therapy through increased apoptosis by inhibition of NF-kappaB.** *Nat Med* 1999, **5:**412–417.
35. Kantidze OL, Kamalyukova IM, Razin SV: **Association of the mammalian transcriptional regulator kaiso with centrosomes and the midbody.** *Cell cycle Georgetown Tex* 2009, **8:**2303–2304.
36. Tong G-X, Yu WM, Beaubier NT, Weeden EM, Hamele-Bena D, Mansukhani MM, O'Toole KM: **Expression of PAX8 in normal and neoplastic renal tissues: an immunohistochemical study.** *Modern pathology an official journal of the United States and Canadian Academy of Pathology Inc* 2009, **22:**1218–1227.
37. Liu H, Hew HC, Lu Z-G, Yamaguchi T, Miki Y, Yoshida K: **DNA damage signalling recruits RREB-1 to the p53 tumour suppressor promoter.** *Biochem J* 2009, **422:**543–551.
38. Ritchie A, Gutierrez O, Fernandez-Luna JL: **PAR bZIP-bik is a novel transcriptional pathway that mediates oxidative stress-induced apoptosis in fibroblasts.** *Cell Death Differ* 2009, **16:**838–846.
39. Frank DA: **STAT3 as a central mediator of neoplastic cellular transformation.** *Cancer Lett* 2007, **251:**199–210.
40. Felty Q, Xiong W-C, Sun D, Sarkar S, Singh KP, Parkash J, Roy D: **Estrogen-induced mitochondrial reactive oxygen species as signal-transducing messengers.** *Biochemistry* 2005, **44:**6900–6909.
41. Kunkle B, Felty Q, Trevino F, Roy D: **Oncomine meta-analysis of breast cancer microarray data identifies upregulation of NRF-1 expression in human breast carcinoma.** *Distribution* 2009:715–719.

42. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic Acids Res* 2005, **33**:3154–3164.
43. Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Mariño-Ramírez L: **The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site.** *Nucleic Acids Res* 2008, **36**:2777–2786.

doi:10.1186/1471-2164-13-199
Cite this article as: Joshi *et al.*: Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics* 2012 **13**:199.

Additional Files to Paper I

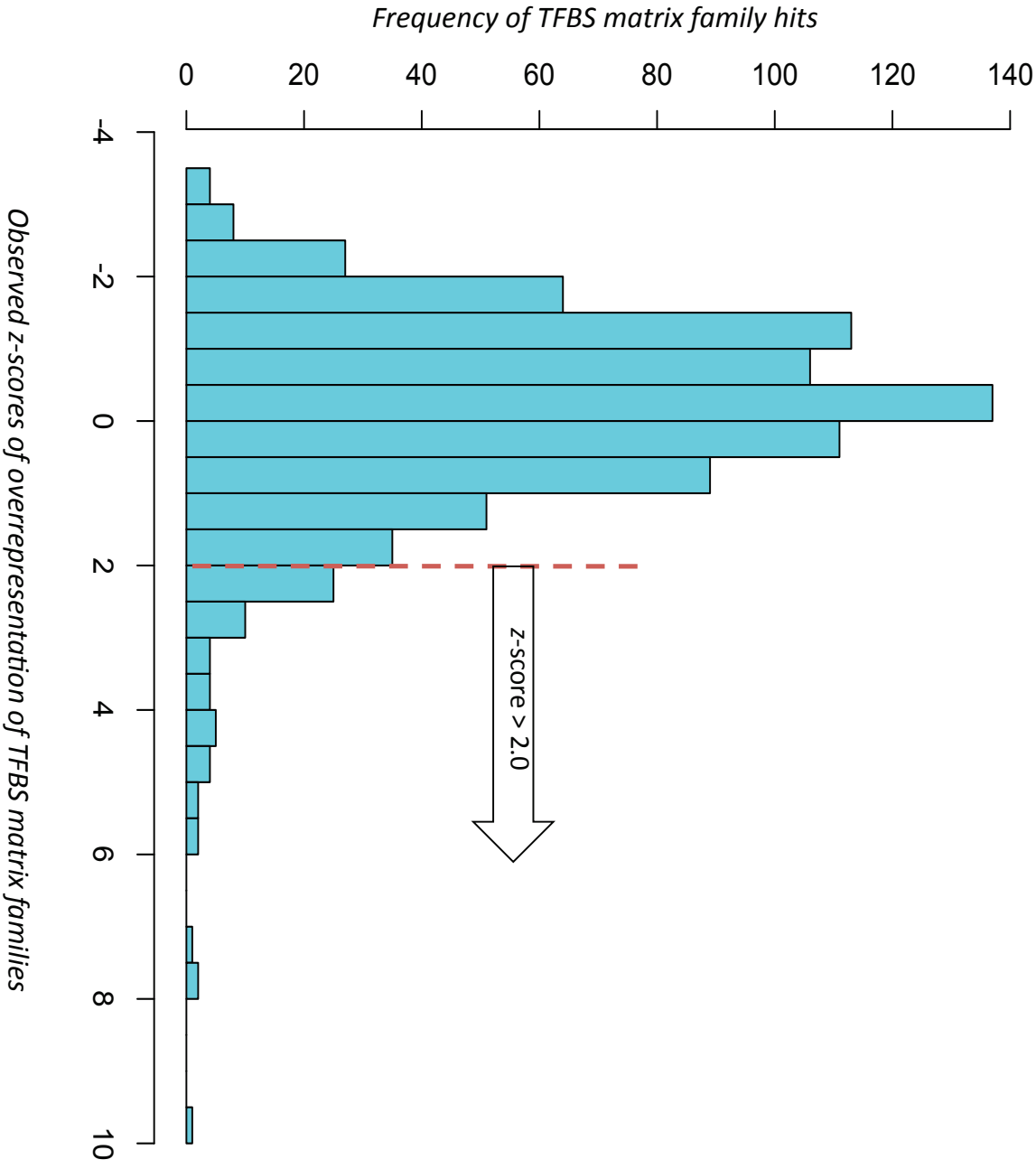
Additional file 1: Table S1

Supplementary Table 1: Table shows the 197 subtype-specific best discriminatory genes, which is a subset of the intrinsic gene-list.

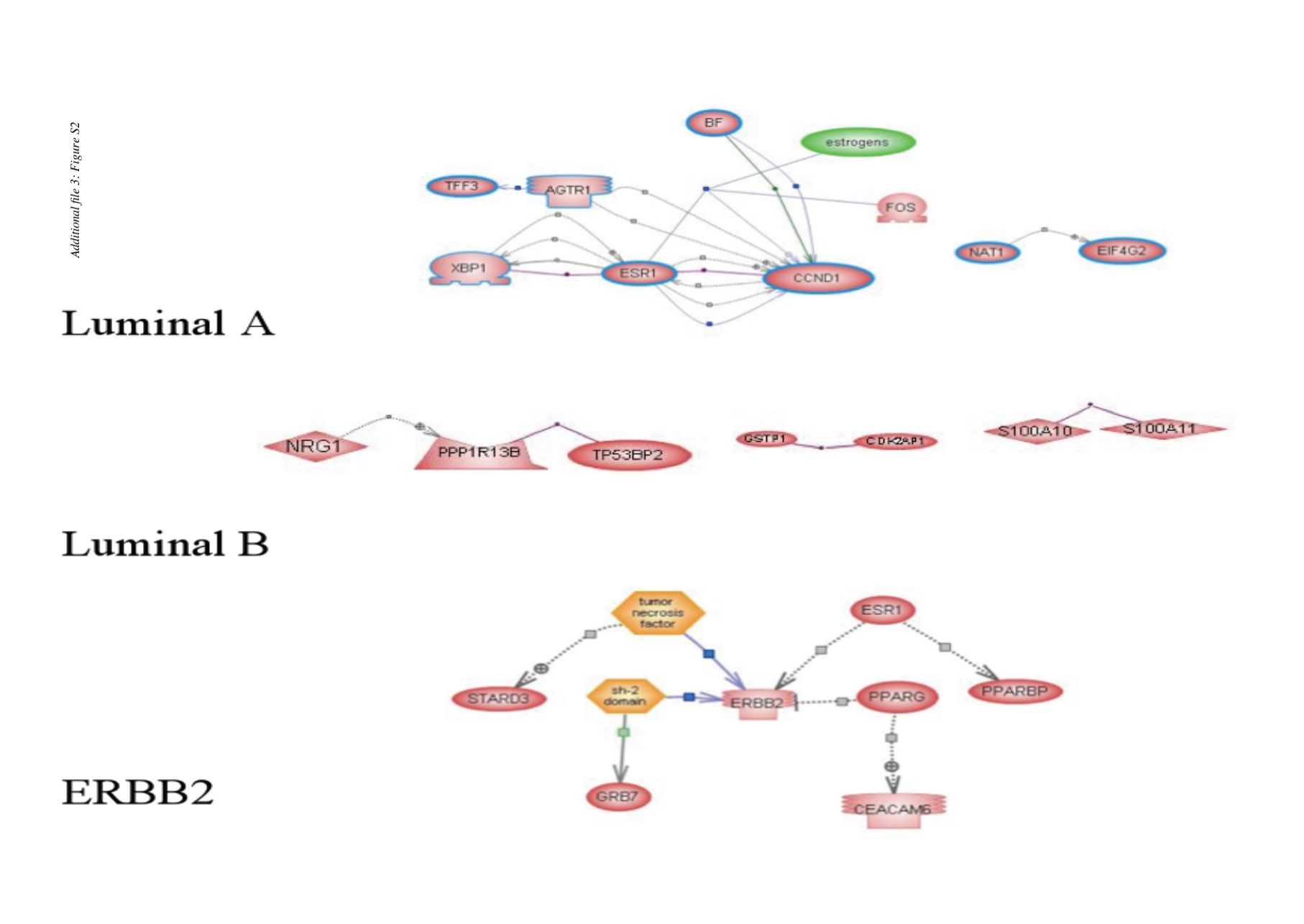
Subtype	Name	Symbol
LUMINAL A	Acyl-Coenzyme A dehydrogenase, short/branched chain	ACADSB
	Adrenergic, alpha-2A-, receptor	ADRA2A
	Angiotensin II receptor, type 1	AGTR1
	Activated leukocyte cell adhesion molecule	ALCAM
	Annexin A9	ANXA9
	N-acylsphingosine amidohydrolase (acid ceramidase) 1	ASAH1
	Beclin 1, autophagy related	BECN1
	Complement factor B	CFB
	Biliverdin reductase A	BLVRA
	Chromosome 14 open reading frame 132	C14orf132
	Complement component 4B (Childo blood group)	C4A
	Calcium/calmodulin-dependent protein kinase II inhibitor 1	CAMK2N1
	Cyclin D1	CCND1
	Cytochrome c oxidase subunit VIc	COX6C
	Carnitine acetyltransferase	CRAT
	Cytochrome b5 type A (microsomal)	CYB5A
	Cytochrome P450, family 2, subfamily A, polypeptide 6	CYP2A6
	Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 2	APPL2
	Receptor accessory protein 5	REEP5
	Ectonucleotide pyrophosphatase/phosphodiesterase 5 (putative function)	ENPP5
	Estrogen receptor 1	ESR1
	Fructose-1,6-bisphosphatase 1	FBP1
	Enoyl Coenzyme A hydratase domain containing 2	ECHDC2
	Acyl-Coenzyme A binding domain containing 4	ACBD4
	Fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	FLT1
	Flavin containing monooxygenase 5	FMO5
	Fibromodulin	FMOD
	Forkhead box A1	FOXA1
	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 10 (GalNAc-T10)	GALNT10
	GATA binding protein 3	GATA3
	Glucocorticoid receptor DNA binding factor 1	GRLF1
	Glutathione S-transferase M3 (brain)	GSTM3
	Hexamethylene bis-acetamide inducible 1	HEXIM1
	Hydroxysteroid (17-beta) dehydrogenase 4	HSD17B4
	KIAA0182	KIAA0182
	PHD finger protein 15	PHF15
	Jumonji domain containing 2B	JMJD2B
	Mediator complex subunit 13-like	MED13L
	Solute carrier family 39 (zinc transporter), member 6	SLC39A6
	Nephronectin	NPNT
	LPS-responsive vesicle trafficking, beach and anchor containing	LRBA
	Basal cell adhesion molecule (Lutheran blood group)	BCAM
	Methylcrotonoyl-Coenzyme A carboxylase 2 (beta)	MCCC2
	Chromosome 10 open reading frame 32	C10orf32
	Sushi, nidogen and EGF-like domains 1	SNED1
	Mahogunin, ring finger 1	MGRN1
	Msh homeobox 2	MSX2
	Mucin 1, cell surface associated	MUC1
	N-acetyltransferase 1 (arylamine N-acetyltransferase)	NAT1
	Transcribed locus	IMAGE:132012
	Neuropeptide Y receptor Y1	NPY1R
	Plasminogen activator, tissue	PLAT
	Sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1	SVEP1

N.B. : Table truncated because of the size. Complete table is available at : http://www.biomedcentral.com/1471-2164/13/199/

Additional file 2:Figure S1.



Additional file 3: Figure S2



Because of the large figure sizes, Additional Files 5–8 (Suppl. Figures S3–S7) are available from
:http://www.biomedcentral.com/1471-2164/13/199/additional

Supplementary Table 2. List of significantly over-represented transcription factor binding site families in subtypes of breast cancers at the cut-off level of z-score >=2.0.

TF Families	Description	Nr of Sequences	Nr of Matches	Expected frequency (genome)	Overrepresentation (genome)	Z-Score (genome)	Expected (promoters)	Overrepresentation (promoters)	Z-Score (promoters)
Basal									
O$TF2B	RNA polymerase II transcription factor II B	6	12	0.82	14.71	11.83	6.35	1.89	2.04
V$BTBF	BTB/POZ (broad complex, TramTrack, Bric-a-brac/pox viruses and zinc fingers) transcription factor	12	14	4.51	3.1	4.23	4.38	3.2	4.36
V$MOKF	Mouse Krueppel like factor	14	27	13.68	1.97	3.47	17.43	1.55	2.17
V$SRFF	Serum response element binding factor	17	35	30.62	1.14	0.7	23.52	1.49	2.27
ERBB2+									
V$EKLF	Basic and erythroid krueppel like factors	10	30	6.85	4.38	8.66	17.35	1.73	2.92
V$GREF	Glucocorticoid responsive and related elements	10	24	13.43	1.79	2.75	12.03	2	3.31
V$MAZF	Myc associated zinc fingers	11	32	4.55	7.03	12.64	16.12	1.98	3.83
V$MZF1	Myeloid zinc finger 1 factors	12	24	7.64	3.14	5.74	13.16	1.82	2.85
V$NFKB	Nuclear factor kappa B/c-rel	11	24	7.29	3.29	6	11.67	2.06	3.46
V$PLAG	Pleomorphic adenoma gene	8	16	2.37	6.76	8.54	8.5	1.88	2.4
V$RREB	Ras-responsive element binding protein	10	13	3.62	3.59	4.67	6.83	1.9	2.17
V$SP1F	GC-Box factors SP1/GC	11	58	9.03	6.43	16.15	28.47	2.04	5.45
V$XBBF	X-box binding factors	8	17	7.25	2.35	3.44	10.02	1.7	2.05
V$ZBPF	Zinc binding protein factors	11	63	8.96	7.03	17.9	32.93	1.91	5.16
Luminal A									
O$INRE	Core promoter initiator elements	39	54	53.92	1	-0.06	40.17	1.34	2.1
O$MTEN	Core promoter motif ten elements	33	75	9.52	7.88	21.07	57.01	1.32	2.32
O$PTBP	Plant TATA binding protein factor	22	110	114.22	0.96	-0.44	54.75	2.01	7.4
O$XCPE	Activator-, mediator- and TBP-dependent core promoter element for RNA polymerase	39	72	9.72	7.4	19.81	55.89	1.29	2.09
V$AP1F	AP1, Activating protein 1	33	58	55.74	1.04	0.24	43.98	1.32	2.04
V$ARID	AT rich interactive domain factor	10	13	6.94	1.87	2.11	4.56	2.85	3.71
V$CLOX	CLOX and CLOX homology (CDP) factors	69	205	284.23	0.72	-4.74	175.47	1.17	2.19
V$EBOX	E-box binding factors	61	123	56.47	2.18	8.79	98.34	1.25	2.44
V$GATA	GATA binding factors	68	168	205.78	0.82	-2.67	135.12	1.24	2.79
V$GLIF	GLI zinc finger family	54	91	37.79	2.41	8.58	71.41	1.27	2.26
V$HAML	Human acute myelogenous leukemia factors	41	51	39.04	1.31	1.83	36.88	1.38	2.24
V$HNF1	Hepatic Nuclear Factor 1	45	94	119.4	0.79	-2.37	71.19	1.32	2.65
V$HOXF	Factors with moderate activity to homeo domain consensus sequence	80	348	482.2	0.72	-6.16	301.07	1.16	2.68

Supplementary Table 3: Table shows the fold over-representation of potential transcriptional factor hits (represented as TFBS families) in subtype-specific gene promoter sequences.

TFBSFamilyName	Basal	ERBB2+	LumA	LumB	Normal-like
O$INRE	1.04	0.69	1.34	0.65	1.23
O$MTEN	1.47	0.85	1.32	1.81	0.87
O$PTBP	0.94	0.51	2.01	0.6	1.07
O$TF2B	1.89	0.34	0.98	2.24	1.32
O$TF2D	0.81	1.32	0.95	1.35	0.86
O$TF3C	0.67	0	0.83	1.36	0
O$VTBP	1.24	0.75	1.03	0.98	1.28
O$XCPE	0.86	1.61	1.29	1.7	1.05
V$AARF	0.84	0.6	0.95	0.99	1.17
V$AHRR	0.72	0.78	0.85	1.71	0.87
V$AIRE	0.75	0.81	1.11	1.05	1.05
V$AP1F	1.24	0.47	1.32	0.52	1.33
V$AP1R	0.93	0.71	0.97	0.75	1.07
V$AP2F	0.87	1.19	1.3	1.21	1.55
V$AP4R	0.6	0.64	1.06	1.12	0.42
V$ARID	0.7	1.52	2.85	0.36	0
V$ATBF	0.66	0.32	1.01	0.79	0.52
V$BARB	0.91	0.49	1.23	0.92	0.95
V$BCL6	0.82	0.95	1.12	1	1.24
V$BNCF	1.08	1.16	0.67	0.46	1.51
V$BPTF	1.24	1.14	0.66	0.54	0.74
V$BRAC	1.39	0.54	1.17	1.21	1.8
V$BRNF	0.86	0	1.08	0.99	0.96
V$BTBF	3.2	1.97	0.64	0.35	0.96
V$CAAT	0.82	0.73	1.18	1.26	1.1
V$CABL	1.02	0.73	0.63	0.87	0.48
V$CART	0.56	0.45	1.08	1.06	0.58
V$CDEF	0.72	0.78	0.75	1.95	0.67
V$CDXF	1.03	0.69	1.1	1.38	1.62
V$CEBP	1.23	0.63	1.07	1.16	0.82
V$CHOP	1.08	1.75	0.59	1.1	0
V$CHRE	0.8	0.69	0.7	1.31	1.12
V$CHRF	0.55	0.98	0.74	0.6	1.66
V$CIZF	0.84	0.9	1.07	0.85	1.32
V$CLOX	1.1	0.99	1.17	1.03	0.77
V$CP2F	1.51	1.37	1.01	0.85	0.67
V$CREB	0.75	1.02	0.98	0.93	0.98
V$CSEN	0.75	2.01	1.34	0.48	1.31
V$CTCF	1.05	1.34	1.15	1.73	1.07
V$DEAF	1.08	1.54	0.82	1.1	0.5
V$DICE	1.07	0.58	1.12	0.95	0.19
V$DMRT	0.86	0.77	1.13	0.99	0.58
V$DMTF	1.54	0.74	1.17	1.04	0.24
V$E2FF	0.88	0.6	0.97	1.87	0.71
V$E4FF	0.64	1.11	1	1.24	1.44
V$EBOX	0.69	0.84	1.25	1.11	0.69
V$EGRF	0.96	1.24	1.06	1.68	0.85
V$EKLF	1.12	1.73	0.99	1.31	0.75
V$EREF	0.87	1.53	0.96	1.01	1.11
V$ETSF	0.86	1.19	0.92	1.14	0.91
V$EVI1	0.71	0.75	0.92	1.09	1.18
V$FAST	0.82	1.23	0.77	0.74	0.97
V$FKHD	0.98	0.61	1.05	0.79	0.82
V$FXRE	0.7	0	1.04	0.81	1.97
V$GABF	0.16	1.59	1.17	0.92	0.11
V$GATA	1.09	0.72	1.24	0.91	0.9
V$GCMF	1.51	1.08	0.6	0.77	0.47
V$GCNR	0.69	0	0.43	0	0
V$GFI1	0.94	1.18	0.97	0.95	0.55
V$GKLF	0.84	1.47	0.88	1.02	0.88
V$GLIF	1.04	1.07	1.27	1.24	0.57
V$GREF	1.04	2	0.78	0.71	0.86
V$GRHL	0.77	0.71	0.65	0.68	0.93
V$GUCE	1.3	0	1.21	1.32	0
V$GZF1	0.64	0.69	1.34	0.57	0.67
V$HAML	1.22	1.5	1.38	1.06	0.73
V$HAND	0.85	1.1	1.08	0.94	0.79
V$HEAT	1.14	1.22	0.96	0.93	0.75
V$HESF	0.73	0.55	1.08	1.27	0.96
V$HICF	0.54	0.58	1.42	1.36	1.12
V$HIFF	0.55	0.88	0.98	1.11	1.24
V$HMTB	1.62	1.28	1.06	0.95	1.67
V$HNF1	0.68	0.39	1.32	0.8	1.07
V$HNF6	0.55	0.6	1.17	0.93	1.08
V$HOMF	0.76	0.73	1	0.84	1.01
V$HOXC	0.74	0.79	0.88	1.07	0.82
V$HOXF	0.84	0.92	1.16	0.9	0.87
V$HOXH	0.69	0.4	1.09	0.96	1.14
V$IKRS	1.09	0.65	1.13	1.05	0.76
V$INSM	1.31	1.56	0.88	1.03	1.52
V$IRFF	1.05	0.82	1.24	1.25	1.12

N.B. : Table truncated because of the size. Complete table is available at : http://www.biomedcentral.com/1471-2164/13/199/

Supplementary Table 4: Table displays the pearson´s correlation between the geometric mean of expression values of transcription factor genes in subtypes and fold overrepresentation of corresponding TFBS familes

Gene Symbol	GeneName	TFBS families	Pearson's correlation
SNFT	Jun dimerization protein p21SNFT	V$AP1F	0.37
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	V$AP1F	0.65
FOSB	FBJ murine osteosarcoma viral oncogene homolog B	V$AP1F	0.83
FOSL2	FOS-like antigen 2	V$AP1F	0.85
FOSL1	FOS-like antigen 1	V$AP1F	0.43
JDP2	jun dimerization protein 2	V$AP1F	0.48
JUN	jun oncogene	V$AP1F	0.74
JUNB	jun B proto-oncogene	V$AP1F	0.72
JUNB	jun B proto-oncogene	V$AP1F	0.66
ARID5B	AT rich interactive domain 5B (MRF1-like)	V$ARID	-0.16
ARID5B	AT rich interactive domain 5B (MRF1-like)	V$ARID	-0.07
ARID5B	AT rich interactive domain 5B (MRF1-like)	V$ARID	0.27
CUTL1	cut-like 1, CCAAT displacement protein (Drosophila)	V$CLOX	0.14
CUTL2	cut-like 2 (Drosophila)	V$CLOX	-0.74
E2F1	E2F transcription factor 1	V$E2FF	0.34
E2F2	E2F transcription factor 2	V$E2FF	0.15
E2F3	E2F transcription factor 3	V$E2FF	-0.21
E2F4	E2F transcription factor 4, p107/p130-binding	V$E2FF	-0.53
E2F4	E2F transcription factor 4, p107/p130-binding	V$E2FF	-0.68
E2F5	E2F transcription factor 5, p130-binding	V$E2FF	0.81
E2F7	E2F transcription factor 7	V$E2FF	0.37
E2F8	E2F transcription factor 8	V$E2FF	0.18
TFDP1	transcription factor Dp-1	V$E2FF	-0.01
TFDP1	transcription factor Dp-1	V$E2FF	-0.05
ATF6	activating transcription factor 6	V$EBOX	0.88
CREBL1	cAMP responsive element binding protein-like 1	V$EBOX	0.49
MAX	MYC associated factor X	V$EBOX	0.85
MGA	MAX gene associated	V$EBOX	-0.58
MLX	MAX-like protein X	V$EBOX	0.26
MLXIPL	MLX interacting protein-like	V$EBOX	0.78
MLXIPL	MLX interacting protein-like	V$EBOX	-0.83
MLXIPL	MLX interacting protein-like	V$EBOX	0.76
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	V$EBOX	-0.38
MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)	V$EBOX	-0.35
TCF4	transcription factor 4	V$EBOX	-0.36
EGR1	early growth response 1	V$EGRF	-0.76
EGR2	early growth response 2 (Krox-20 homolog, Drosophila)	V$EGRF	-0.78
EGR3	early growth response 3	V$EGRF	-0.51
ZBTB7A	zinc finger and BTB domain containing 7A	V$EGRF	0.29
ZBTB7B	zinc finger and BTB domain containing 7B	V$EGRF	-0.02
KLF2	Kruppel-like factor 2 (lung)	V$EKLF	-0.01
KLF2	Kruppel-like factor 2 (lung)	V$EKLF	-0.21
KLF3	Kruppel-like factor 3 (basic)	V$EKLF	-0.16
KLF4	Kruppel-like factor 4 (gut)	V$EKLF	-0.13
KLF6	Kruppel-like factor 6	V$EKLF	-0.34
KLF6	Kruppel-like factor 6	V$EKLF	-0.75
KLF6	Kruppel-like factor 6	V$EKLF	-0.21
KLF7	Kruppel-like factor 7 (ubiquitous)	V$EKLF	0.86
KLF8	Kruppel-like factor 8	V$EKLF	0.09
KLF8	Kruppel-like factor 8	V$EKLF	-0.21
TRPS1	trichorhinophalangeal syndrome I	V$GATA	0.37
GATA2	GATA binding protein 2	V$GATA	-0.04
GATA3	GATA binding protein 3	V$GATA	0.35
GATA6	GATA binding protein 6	V$GATA	-0.67
GATAD2A	GATA zinc finger domain containing 2A	V$GATA	-0.31
GATAD1	GATA zinc finger domain containing 1	V$GATA	0.61
GATAD1	GATA zinc finger domain containing 1	V$GATA	0.50
GLI1	glioma-associated oncogene homolog 1 (zinc finger protein)	V$GLIF	-0.87
GLI2	GLI-Kruppel family member GLI2	V$GLIF	-0.84
GLI2	GLI-Kruppel family member GLI2	V$GLIF	-0.79
GLI3	GLI-Kruppel family member GLI3 (Greig cephalopolysyndactyly syndrome)	V$GLIF	-0.10
GLIS1	GLIS family zinc finger 1	V$GLIF	-0.69
GLIS2	GLIS family zinc finger 2	V$GLIF	0.17
ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)	V$GLIF	-0.18
ZIC1	Zic family member 1 (odd-paired homolog, Drosophila)	V$GLIF	-0.22
ZIC4	Zic family member 4	V$GLIF	-0.02
ZIC5	Zic family member 5 (odd-paired homolog, Drosophila)	V$GLIF	-0.22
AR	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)	V$GREF	-0.09
NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	V$GREF	-0.48

N.B. : Table truncated because of the size. Complete table is available at : http://www.biomedcentral.com/1471-2164/13/199/additional

II

British Journal of Cancer (2012) 107, 1722–1728

www.bjcancer.com

Potential tumorigenic programs associated with *TP53* mutation status reveal role of VEGF pathway

H Joshi*[,1], G Bhanot[2,3], A-L Børresen-Dale[4,5] and V Kristensen[1,4,5]

[1]Medical Division (EpiGen), Akershus University Hospital and University of Oslo, Lorenskog 1478, Norway; [2]Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, USA; [3]Department of Physics and BioMaPS Institute, Rutgers University, Piscataway, NJ 08854, USA; [4]Institute of Clinical Medicine, University of Oslo, Oslo 0450, Norway; [5]Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo 0310, Norway

BACKGROUND: Targeting differentially activated or perturbed tumour pathways is the key idea in individualised cancer therapy, which is emerging as an important option in treating cancers with poor prognostic profiles. *TP53* mutation status is known as a core determinant of survival in breast cancer. The pathways disrupted in association with *TP53* mutation status in tumours are not well characterised.
METHOD: In this study, we stratify breast cancers based on their *TP53* mutation status and identify the set of dysregulated tumorigenic pathways and corresponding candidate driver genes using breast cancer gene expression profiles. Expressions of these genes were evaluated for their effect on patient survival first in univariate models, followed by multivariate models with TP53 status as a covariate.
RESULTS: The most strongly differentially enriched pathways between breast cancers stratified by *TP53* mutation status include in addition to TP53 signalling, several known cancer pathways involved in renal, prostate, pancreatic, colorectal, lung and other cancers, and signalling pathways such as calcium signalling, MAPK, ERBB and vascular endothelial growth factor (VEGF) signalling pathways. We found that mutant TP53 in conjunction with active estrogen receptor (ER) signalling significantly influence survival. We also found that upregulation of *VEGFA* mRNA levels in association with active ER signalling is a significant marker for poor survival, even in the presence of wild-type TP53.
CONCLUSION: Mutation status of *TP53* in breast cancer involves wide ranging derangement of several pathways. Among the candidate genes of the significantly deranged pathways, we identified *VEGFA* expression as an important marker of survival even when controlled by *TP53* mutation status. Interestingly, independent of the *TP53* mutation status, the survival effect of *VEGFA* was found significant in patients with active ER signalling (ER/PgR +), but not in those with ER/PgR − status. Therefore, we propose more studies to focus on the role of complex interplay between TP53, ER and VEGF signalling from therapeutic and prognostic context in breast cancer.
British Journal of Cancer (2012) **107,** 1722–1728. doi:10.1038/bjc.2012.461 www.bjcancer.com
Published online 18 October 2012

Keywords: breast cancer; *TP53* mutation status; estrogen receptor signalling; vascular endothelial growth factor signalling; dysregulated pathways; survival

The fact that nearly 30% of early-diagnosed breast cancer cases might eventually develop recurrent or metastatic disease (O'Shaughnessy, 2005) – underscores the priority to explore the mechanisms of advanced disease. The TP53 protein is an important clinical biomarker of breast cancer because of its association with tumour progression (Norberg *et al*, 2001), metastatic potential (D'Assoro *et al*, 2010), early relapse (Aas *et al*, 1996), response to chemotherapy (Aas *et al*, 1996; Kandioler-Eckersberger *et al*, 2000; Bertheau *et al*, 2007), and ultimately, to prognosis and survival (Børresen *et al*, 1995; Berns *et al*, 2000; Olivier *et al*, 2006). It is also of relevance to molecular subtypes of breast cancer (Miller *et al*, 2005; Langerød *et al*, 2007). Whereas ~70% of breast cancers with wild-type *TP53* are mostly of the Luminal A subtype, mutant *TP53* is common in the remaining 30%, which have a poorer prognosis and are classified as triple negative or luminal B. The focus of this work is to identify diagnostic, prognostic and therapeutic biomarkers associated with pathways perturbed by *TP53* mutations and understand their relationship to patient survival in breast cancer, under current therapeutic protocols.

TP53 is a key regulator of programmed cell death, cell cycle, DNA repair and genomic stability. In response to stimulus-specific post-transcriptional modification, TP53 regulates genes, which activate specific cellular programs. The TP53 protein has three major functional domains: a transactivation domain at its N-terminal, a central DNA-binding domain (which includes mutation hotspots) and tetramerization and regulatory domains at the C-terminal. The location and type of *TP53* mutation affect the ability of TP53 to regulate its target genes, leading to aberrant functions (Blandino *et al*, 1999) with clinical implications (Kim and Deppert, 2006). Characterisation of the differential activation of key pathways and candidate genes according to the TP53 mutation status may therefore identify mechanisms correlated with *TP53* mutation status in breast cancer.

In this study, we stratify breast cancers based on their *TP53* mutation status and identify the set of dysregulated tumorigenic pathways and their candidate driver genes by using gene

*Correspondence: Dr H Joshi; E-mail: Dr.Himanshu.Joshi@gmail.com
Received 29 May 2012; revised 21 August 2012; accepted 18 September 2012; published online 18 October 2012

expression data sets obtained from tumours. The goal is to infer the class-specific candidate gene signature by identifying weak to moderate, but coherent gene expressions that significantly influence tumorigenic pathways and survival.

RESULTS

We first categorised breast cancer samples by their corresponding *TP53* mutation status, as described in Supplementary Table 1 and performed analysis as shown in the flow-chart (Supplementary Figure 1).

Candidate driver pathways differentially perturbed by *TP53* mutations

Enrichment analysis of pathways between mutation status classes was performed using globaltest (Goeman *et al*, 2011) and *SAM-GS* (Dinu *et al*, 2007) on the primary and combined validation data set. Globaltest, although being sensitive to genes with smaller regression coefficients, its results might be influenced by the standardisation and normalisation procedures. SAM-GS on the other hand is shown to have relatively higher power in the lower alpha-level region, thus can better focus on pathways of greatest interest (Liu *et al*, 2007). Therefore, we use a combination of the two approaches here. The list of differentially enriched KEGG (Kanehisa and Goto, 2000) pathways identified by each of the methods on each of the data set is shown together in Supplementary Table 2. A set of 40 pathways inferred as commonly significant by both the methods in both data sets (Table 1) – are graphically presented as an enrichment map color-coded according to globaltest FDR corrected *P*-values (Supplementary Figure 2). The most dysregulated pathways included a group of key signalling pathways – such as p53 signalling, calcium signalling, MAPK, ErbB, vascular endothelial growth factor (VEGF) signalling and various cancer pathways.

Table 1 Consensus list of differentially enriched pathways between two *TP53* mutation status classes (wild-type *TP53* profiles compared with the mutant *TP53* profiles), based on pathway analysis performed by using two approaches – globaltest and SAM-GS on primary ($n = 111$ samples) and validation data sets (a combined cross-platform data set with $n = 327$)

		Primary data set		Validation data set	
KEGGID	**KEGG pathway name**	**Asymptotic global test BH corrected *P*-value**	**SAM-GS FDR adj *P*-value**	**Asymptotic global test BH corrected *P*-value**	**SAM-GS FDR adj *P*-value**
hsa:00230	Purine metabolism	1.8E − 09	< 10e − 6	2.37E − 36	< 10e − 6
hsa:04115	p53-signalling pathway	1.8E − 09	< 10e − 6	2.43E − 34	< 10e − 6
hsa:05211	Renal cell carcinoma	3.72E − 09	< 10e − 6	1.32E − 17	< 10e − 6
hsa:05200	Pathways in cancer	1.1E − 08	< 10e − 6	6.86E − 29	< 10e − 6
hsa:05215	Prostate cancer	1.1E − 08	< 10e − 6	1.32E − 29	< 10e − 6
hsa:04020	Calcium-signalling pathway	4.12E − 08	< 10e − 6	4.81E − 27	< 10e − 6
hsa:00260	Glycine, serine and threonine metabolism	4.73E − 08	< 10e − 6	1.44E − 25	< 10e − 6
hsa:05212	Pancreatic cancer	5.65E − 08	< 10e − 6	1.15E − 39	< 10e − 6
hsa:04340	Hedgehog-signalling pathway	6.02E − 08	< 10e − 6	5.76E − 21	< 10e − 6
hsa:05222	Small-cell lung cancer	7.93E − 08	< 10e − 6	6.75E − 40	< 10e − 6
hsa:04120	Ubiquitin-mediated proteolysis	0.00000012	< 10e − 6	3.26E − 40	< 10e − 6
hsa:04910	Insulin signalling pathway	0.00000012	< 10e − 6	5.83E − 27	< 10e − 6
hsa:00051	Fructose and mannose metabolism	1.28E − 07	< 10e − 6	2.51E − 30	< 10e − 6
hsa:05218	Melanoma	0.00000014	< 10e − 6	1.7E − 17	< 10e − 6
hsa:04150	mTOR-signalling pathway	1.68E − 07	< 10e − 6	9.66E − 26	< 10e − 6
hsa:00380	Tryptophan metabolism	1.96E − 07	< 10e − 6	1.48E − 08	< 10e − 6
hsa:04144	Endocytosis	2.39E − 07	< 10e − 6	4.96E − 24	< 10e − 6
hsa:00330	Arginine and proline metabolism	0.00000025	< 10e − 6	1.29E − 18	< 10e − 6
hsa:05214	Glioma	0.00000025	< 10e − 6	1.47E − 14	< 10e − 6
hsa:04010	MAPK-signalling pathway	0.00000031	< 10e − 6	2.44E − 34	< 10e − 6
hsa:04012	ErbB-signalling pathway	3.65E − 07	< 10e − 6	2.68E − 17	< 10e − 6
hsa:04520	Adherens junction	4.03E − 07	< 10e − 6	9.78E − 13	< 10e − 6
hsa:05217	Basal cell carcinoma	0.00000048	< 10e − 6	6.47E − 11	< 10e − 6
hsa:00600	Sphingolipid metabolism	4.94E − 07	< 10e − 6	4.67E − 14	< 10e − 6
hsa:05120	Epithelial cell signalling in *Helicobacter pylori* infection	5.79E − 07	< 10e − 6	1.45E − 11	< 10e − 6
hsa:04722	Neurotrophin-signalling pathway	6.72E − 07	< 10e − 6	1.09E − 21	< 10e − 6
hsa:04912	GnRH-signalling pathway	8.22E − 07	< 10e − 6	6E − 18	< 10e − 6
hsa:05219	Bladder cancer	8.23E − 07	< 10e − 6	1.61E − 17	< 10e − 6
hsa:05210	Colorectal cancer	0.00000116	< 10e − 6	3.9E − 11	< 10e − 6
hsa:04070	Phosphatidylinositol-signalling system	0.00000117	< 10e − 6	2.16E − 12	< 10e − 6
hsa:04110	Cell cycle	0.00000125	< 10e − 6	3.7E − 27	< 10e − 6
hsa:04370	VEGF-signalling pathway	0.00000153	< 10e − 6	1.01E − 07	< 10e − 6
hsa:05221	Acute myeloid leukaemia	0.00000205	< 10e − 6	6.36E − 12	< 10e − 6
hsa:00270	Cysteine and methionine metabolism	0.0000036	< 10e − 6	1.24E − 25	< 10e − 6
hsa:04530	Tight junction	0.00000531	< 10e − 6	6.85E − 18	< 10e − 6
hsa:04350	TGF-β-signalling pathway	0.00000725	< 10e − 6	5.93E − 14	< 10e − 6
hsa:04310	Wnt-signalling pathway	0.0000103	< 10e − 6	8.24E − 19	< 10e − 6
hsa:00590	Arachidonic acid metabolism	0.0000146	< 10e − 6	1.16E − 11	< 10e − 6
hsa:05213	Endometrial cancer	0.000018	< 10e − 6	0.00000131	< 10e − 6
hsa:04142	Lysosome	0.0000489	< 10e − 6	6.09E − 18	< 10e − 6

Abbreviations: BH, Benjamini-Hochberg; FDR, false discovery rate; MAPK, mitogen-activated protein kinase; mTOR, mammalian target of rapamycin; SAM-GS, significance analysis of microarrays for genesets; TGF, tumour growth factor; VEGF, vascular endothelial growth factor. The full pathway lists that show significance of differential enrichment in each individual data set are shown with their respective *P*-values of significance in Supplementary Table 2.

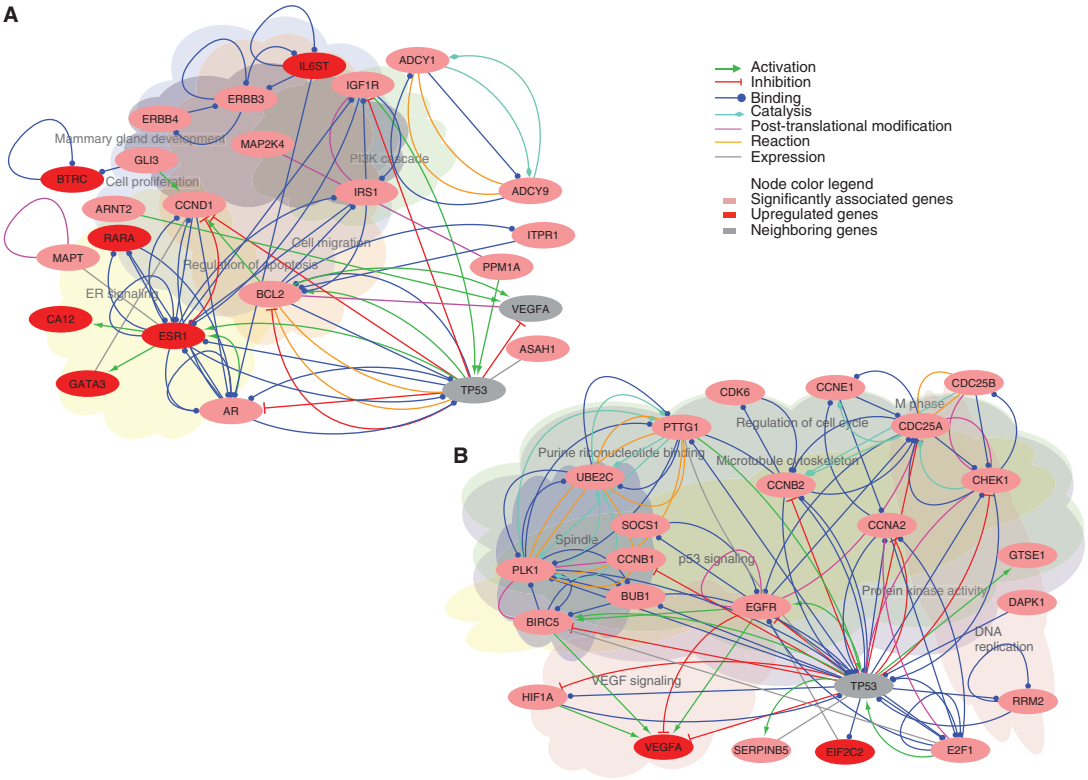


Figure 1 *TP53* mutation status-specific network of potential candidate driver genes shown based on their known and predicted functional interactions. (**A**) Network for wild-type *TP53* breast cancer profiles. (**B**) Network for mutant *TP53* breast cancer profiles. Significant association of gene means significant non-zero regression coefficient of a gene in a significantly differentially enriched KEGG pathway. Gene upregulation means its class-specific upward biased expression pattern, inferred by the rank-sum statistic of the modified Kolmogorov–Smirnov test. Relevant biological processes represented by these genes are also highlighted in background.

Candidate genes deregulated according to the *TP53* mutation status

Candidate genes were identified by applying a combination of two mutually complementary approaches: pathway-based gene-search that infers class-specific association (globaltest) and pathway-independent search that identifies individual genes with class-specific upregulation (modified Kolmogorov-Smirnov approach) on both primary and validation data sets (Supplementary Tables 3 and 4). Combining genesets inferred by these two approaches would help to account for the genes with smaller as well as larger effects on the overall biological condition. A consensus genelist (Supplementary Table 5) of 112 genes consists of genes inferred as significant at least by either of the two statistical tests (but not necessarily by the same test) in both primary and validation data sets, as shown in the Venn diagram (Supplementary Figure 3). Class-specific predicted functional networks based on these genesets are plotted in Figures 1A and B for wild-type and mutant *TP53* samples, respectively. These networks reflect the key genes and corresponding processes that have potential functional implication in association with the one of the *TP53* mutation status class. Wild-type *TP53* samples showed significance of genes involved in estrogen receptor (ER) signalling, whereas mutant *TP53* samples in proliferative processes. Besides, GO terms–response to insulin stimulus and mammary gland development in wild-type and protein kinase activity, mitotic cell cycle, microtubule cytoskeleton in mutant *TP53* class were over-represented (Supplementary Figure 4).

Association of EMT and stemness to *TP53* mutation status

Aberrant TP53 function is shown to induce epithelial-mesenchymal transition (EMT) and thereby confers stemness properties to the cancer cells (Dhar *et al*, 2008). Therefore, we compared our inferred *TP53* status-specific candidate genesets with the published EMT and stemness marker sets. We found that mutant *TP53*-marker geneset was significantly associated with embryonic stem cell (ESC) and its TP53 targets (p53ESC) genesets (P-value <0.05). Whereas wild-type *TP53* signature was found significantly associated with PRC2 targets (P-value: 0.003) (Table 2). Top 1000 upregulated genes (according the signal-to-noise ratio) in mutant *TP53* class were significantly associated with EMT, ESC and induced pluripotent stem cell marker genesets. Moreover, KEGG pathways involved in stemness and EMT properties such as TGFβ, wnt signalling were found differentially enriched (Supplementary Table 6b).

Vascular endothelial growth factor A upregulation with wild-type *TP53* associates with activation of pro-angiogenic and pro-metastatic biological processes

Among the inferred candidate genes that were found upregulated and/or significantly associated to one of the *TP53* mutation status class, 47 genes showed univariate significance to overall patient survival. Vascular endothelial growth factor A (*VEGFA*)

Table 2 Association between the inferred *TP53* mutation status-specific signatures with previously reported EMT and stemness markers. Statistical significance of differential expressed geneset overlapping the stemness and epithelial-mesenchymal transition (EMT) marker genelists[a]. Statistical significance was computed by applying hypergeometric test[b]

	wt*TP53* signature		Mutant *TP53* signature		Top 1000 genes ranked acc to absolute SNR (wt *vs* mt*TP53* BC)		Top 1000 mt*TP53*-upregulated genes ranked acc to SNR	
EMT and stemness geneset and its transcript size	**Number of overlapping genes**	***P*-value**	**Number of overlapping genes**	***P*-value**	**Number of overlapping genes**	***P*-value**	**Number of overlapping genes**	***P*-value**
EMT ($n=497$)	0	NS	1	NS	11	NS	15	0.031
ESC ($n=553$)	0	NS	14	2.65E − 13	22	2.60E − 04	35	4.34E − 11
PRC2 ($n=1016$)	7	3.25E − 03	0	NS	25	NS	19	NS
iPSC ($n=597$)	1	NS	3	NS	17	4.50E − 02	22	1.50E − 03
p53esc ($n=912$)	2	NS	5	2.66E − 02	12	NS	15	NS

Abbreviations: BC, breast cancer; ESC, embryonic stem cell; iPSC, induced pluripotent stem cell; NS, not significant; p53esc, p53 targets identified in murine embryonic stem cells; PRC2, polycomb repressive complex 2; SNR, signal-to-noise ratio. [a]Sources of the genelists are described in the Supplementary Table 6A. [b]Statistical significance was evaluated by Fisher's exact test, in instances where number of overlapping genes $\leqslant 5$.

maintained significance in multivariate model (Supplementary Table 7), even after adjusting for *TP53* mutation status. *VEGFA* might be induced by estrogen receptor in breast cancer cells (Buteau-Lozano *et al*, 2002; Applanat *et al*, 2008). Besides, wild-type TP53 could block VEGFA function induced by active estrogen receptor signalling (Liang *et al*, 2005). However, implications of *VEGFA* in wild-type *TP53*/ER+ patients are less understood. We therefore analysed this subgroup separately by using the globaltest and moderated *t-test* (Smyth, 2004).

Using moderated *t-test* of differential expression on a cross-platform compiled data set, we found 516 gene features (Supplementary Table 8a) differentially expressed between *VEGFA* upregulation (*VEGFA*+) *vs VEGFA* normal/− samples (*VEGFA*−/N). *IGF1* and *PPARG* were found differentially down-regulated in samples with *VEGFA* upregulation. A GO analysis identified pathways associated with blood vessel morphogenesis, cell migration and regulation of VEGF signalling pathway. The complete list of over-represented GO terms and predicted functional interactions are shown in Supplementary Table 9 and Supplementary Figure 5, respectively. Notably, *VEGFA*+ *vs VEGFA*−/N comparison for mutant *TP53* subgroup does not show any remarkable difference apart from differential expression of *VEGFA* itself and pH regulator *CA9* (Supplementary Table 8b).

Tumours overexpressing *VEGFA* (both ER+ wild-type *TP53* and mutant *TP53* irrespective of ER status) show a differential enrichment of the mTOR-signalling pathway compared with normal/downregulated *VEGFA* samples. *VEGFA*+/ER+ wild-type *TP53* samples showed significant association of *EIF4EBP1, MAPK1* (*P*-value <0.05) and weak association of *MTOR, ULK3* and *RPTOR*. Conversely, *PIK3CA* and *IGF1* were significantly associated with *VEGFA* N/− tumours (Supplementary Figure 5 and Supplementary Table 10a). Interestingly, different sets of genes, although involved in the same pathways were found associated with *VEGFA* status in the mutant *TP53* subgroup (Supplementary Table 10b).

TP53 mutation, ER status and *VEGFA* upregulation influence survival

Samples were substratified according to the ER status in each *TP53* mutation class. While comparing the ER+/mutant *TP53* to the ER+/wild-type *TP53* samples, we noted a death hazard ratio (HR) of 2.15 (95% CI: 1.25–3.70) and likelihood *P*-value <0.01. On the other hand, ER− samples showed weaker significance ($P=0.2$; HR: 2.6; 95% CI: 1.14–5.91). As progesterone receptor (PgR) positivity is a better marker of active ER signalling (Bardou *et al*, 2003), we also used PgR status as an indicator of active ER signalling. PgR+ samples showed a significant survival difference between mutant and wild-type tumours ($P=1.53e-05$, HR: 7.2, 95% CI: 3.03–17.1). However, PgR-tumours do not show significant survival differences ($P>0.1$) (Figure 2A and B). On the basis of these findings, we propose that active ER signalling can influence the effect of mutant TP53 on survival.

As *VEGFA* expression is observed here as a significant influencer on survival even after controlling for *TP53* status, we reanalyzed the above effects by adding *VEGFA* expression status as a covariate. Among ER+ group, the overall patient survival was significantly influenced by *TP53* mutation status and *VEGFA* (model significance $=0.0005$) with their corresponding HR $=2.02$ and 2.08, compared with baseline risk for wild-type *TP53* and *VEGFA* normal/downregulation. Even stronger effect was observed after excluding samples with non-missense mutant *TP53* (*P*-value $=0.0001$, HR $=2.38$ and 2.11, respectively). Survival effect of *TP53* mutation status and *VEGFA* was stronger in PgR+ cases (HR $=2.35$, 95% CI: 1.17–4.74 for *VEGFA* upregulation, HR $=5.2$, 95% CI: 2.43–11.1 for mutant *TP53* status, and overall likelihood ratio test $P=2.76e-6$), but non-significant effect in PgR−cases (Figure 2C and D). Although active ER signalling in general is known to predict better prognosis, these findings show that irrespective of the TP53 mutation status, ER+ cases with high mRNA levels of *VEGFA* indicates poor prognosis. Interestingly, despite of the lowest occurrence of cases with upregulated *VEGFA* in ER+/wtTP53 subgroup (Supplementary Figure 6), its prognostic significance underscores further exploration.

DISCUSSION

Our findings show predominance of ER signalling in breast cancers with wild-type *TP53,* marked by the upregulation of *ESR1*, GATA-binding protein 3, retinoic acid receptor alpha (*RARα*) and *CA12*. Estrogen receptor α, a direct transcriptional activator of RARα (Han *et al*, 1997), mediates anti-proliferative response by vitamin A metabolite (all-*trans*-retinoic acid) in breast cancer cells (Dawson *et al*, 1995). Retinoic acid receptor α is a rate-limiting factor for ER transcriptional activity (Ross-Innes *et al*, 2010). Co-expression of *BCL2, ERBB4, IGF1R, IRS1* was also found in this group. Our observation of consistent upregulation of *CA12, AGR3, IL6ST* and *STC2* genes is in agreement with their previously reported association with ER+ breast cancers. Our findings also showed upregulation of *SIRT3*, a mitochondrial p53 activity regulator, necessary for averting TP53-mediated growth arrest (Li *et al*, 2010). Predicted functional network (Figure 1A) provides a hint that genes involved in ER signalling form a core group of

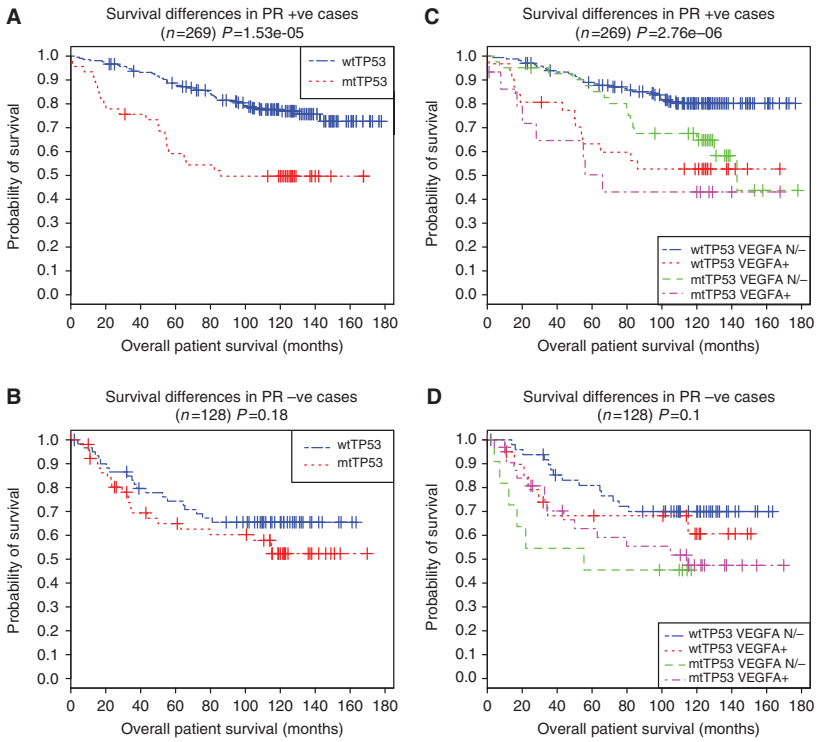


Figure 2 Overall patient survival differs significantly according to the *TP53* mutation status and *VEGFA* expression status in PgR+ and PgR− subgroups of patients. Survival differences between wild-type *TP53* and mutant *TP53* in each of the subgroups are shown in Kaplan–Meier plots shown in **A** and **B**. Survival differences of four classes: (1) wild-type *TP53* and *VEGFA* normal/downregulation (wt*TP53 VEGFA* N/−); (2) wild-type *TP53* and *VEGFA* upregulation (wt*TP53 VEGFA*+); (3) mutant *TP53* and *VEGFA* normal/downregulation (mt*TP53 VEGFA* N/−); and (4) mutant *TP53* and *VEGFA* upregulation (mt*TP53 VEGFA*+) – in PgR+ and PgR− subgroups are shown in **C** and **D**. Significance of overall model is based on the likelihood ratio test *P*-value.

interactions in *TP53* wild-type tumours. A strong relationship between ER signalling and TP53 can be observed in our results. This relationship also has got implications on proliferation and treatment responsiveness. The presence of wild-type TP53 improves sensitivity to Tamoxifen (Berns *et al*, 2000) and inhibits ER cross-talk with the EGFR/HER2 pathways (Fernandez-Cuesta *et al*, 2010). Experimental observations have provided evidence about potential direct ER–TP53 interactions (Liu *et al*, 2006). However, these complex interactions and their effects on transactivation activity of TP53 and ERα in ER+ breast cancer remains to be understood. Given that *TP53* status is an important predictor of response in patients receiving therapy targeting the ER pathway (SERM), we expect that TP53 retains a subset of functions necessary for the response to such therapy.

Genes in pathways related to cell cycle, angiogenesis, chromosomal instability and metastasis were significantly affected in mutant *TP53* tumours. We found the gene *BUB1* and spindle-checkpoint associated kinases were significantly associated with *TP53* mutant tumours. In the presence of dysfunctional TP53, their aberrant expression can cause genomic instability, leading to aneuploidy and malignant transformation (Gjoerup *et al*, 2007). Other genes associated with mutant *TP53* included ones involved in proliferation, angiogenesis and metastasis-*VEGFA, HIF1α, E2F1, CDK6* and *EGFR*.

VEGFA upregulation is an important indicator of pro-angiogenic and pro-metastatic activity. Dysregulation of TP53-VEGF signalling may potentially be a key event in breast cancers with mutant *TP53*. Mutant *TP53* may facilitate this tumorigenic programme by: passing the direct survival advantage to malignant cells, by facilitating the VEGF-mediated enhanced cell migration, angiogenesis and metastasis or by overcoming the regulation by *ETS1* (Dittmer, 2003). Active ER signalling and mutant TP53 are also reported to activate VEGF and mark poor prognosis (Berns *et al*, 2003). In our data, we see that mutant *TP53* and VEGF upregulation significantly affects patient survival in ER+/PgR+ samples, but not in ER−/PgR− samples. Activation of VEGFA may also be attributed to the expression of EGFR (Maity *et al*, 2000) or CDK6, which can correlate with the expression of mutant TP53 (Wyllie *et al*, 2003) and potentially delay cell senescence. Thus, besides the direct effects of lost TP53 function, other related opportunistic mechanisms, such as dysregulated proliferative effects of VEGFA may contribute the overall manifestation.

ER+/wild-type *TP53* samples showed relatively low occurrence of *VEGFA* upregulation but poor survival profile. ER-mediated induction of VEGF (Berns *et al*, 2003; Applanat *et al*, 2008) and VEGF regulation by TP53 (Liang *et al*, 2005) suggests a complex interplay between these three signalling mechanisms. This group also showed the differential enrichment of mTOR signalling. Co-activation of VEGF and mTOR pathway components has been previously reported (Trinh *et al*, 2009). Thus, VEGFA may represent a biomarker of interest to identify the target subset of ER+ breast cancer patients who might benefit from early administration of VEGFA or mTOR-targeted therapy.

MATERIALS AND METHODS

Agilent chip based gene expression data for a subset of 111 breast cancer cases from (Enerly *et al*, 2011) GEO (accession number GSE19783) was used as the primary data set. *TP53* mutations for the primary data in coding regions of exons 2–11 and clinical data were obtained from (Naume *et al*, 2007). Expression data used for validation was obtained from GEO (accession number GSE3494) and from Stanford Microarray Database. Clinical and *TP53* data for these data sets were obtained from (Miller *et al*, 2005; Langerød *et al*, 2007).

Methods used to merge data sets to form a validation data set

Two expression data sets (Miller *et al*, 2005; Langerød *et al*, 2007) from independent studies and different technology platform were preprocessed, quantile normalised and combined based on UniGene identifiers. Batch effects were corrected by applying parametric empirical Bayes method (Johnson *et al*, 2007).

Differential enrichment of pathways and candidate genes

The globaltest (Goeman *et al*, 2011) uses a regression model where genes are covariates and sample classes are response variables. Significant association of gene means significant non-zero regression coefficient of a gene in a geneset (here a particular KEGG pathway). SAM-GS is another geneset enrichment analysis method based on the *t*-like statistic for assessing the permutation-based significance of association between an individual pathway and a phenotype of interest. KEGG pathways inferred as significant by globaltest at FDR corrected *P*-value of $10e-5$ and validated by SAM-GS (Dinu *et al*, 2007) at FDR corrected *P*-value cutoff = $10e-6$ on both primary and validation data sets were analysed by *post-hoc* covariate test to identify significant genes. Gene upregulation means its class-specific upward biased expression pattern, inferred by the rank-sum statistic of the modified Kolmogorov–Smirnov test (Yang *et al*, 2010).

Class-specific predicted functional interactions between genes in the genesets were obtained from STRING database (Jensen *et al*, 2009).

Pathways enrichment and GO analysis

Gene Ontology (GO) analysis was performed for each *TP53* mutation status-specific genesets using DAVID (Huang *et al*, 2009) by Fisher's exact test with human whole genome as a background. Differentially enriched pathways and GO terms were graphically presented as Enrichment map (Merico, 2009), with nodes color-coded by FDR-adjusted *P*-value of significance and node-size proportionate to number of genes in the pathway. Fraction of overlapping genes between any two pathways is represented by the edge thickness, with cutoff overlap coefficient of 0.1.

Association of TP53 biology with EMT and stemness marker signatures

Inferred class-specific genesets were tested by hypergeometric test for their association to the published EMT and stemness marker genesets shown in Supplementary Table 6a. A larger genelist inferred by using signal-to-noise ratio between *TP53* mutation status classes was also tested for its association to these published genesets.

Survival analysis

A combined cohort of 438 cases obtained by merging clinical data from three individual clinical data sets (Supplementary Table 1) was used. Kaplan–Meier estimation of survival and computation of Cox proportional hazards frailty model for the death event was performed by using R package *survival* (Therneau and Lumley, 2009). Inferred candidate genes were assessed for their uni-/multivariate effect on survival. The effect of TP53 mutation status together with genes that maintain significance in a multivariate model (*VEGFA* expression status) and predicted subtype (Parker *et al*, 2009)- was computed with and without stratification by ER/PgR status.

Discretisation of gene expression

The mRNA expression levels of candidate genes were discretised into two levels using mean (μ) + 0.5*standard deviation (s.d.) as a cutoff in each data set.

Analysis was performed by using R (R Development Core Team, 2011).

ACKNOWLEDGEMENTS

This paper is a part of the doctoral thesis work of HJ. His work was supported by grant number 2789119 from Helse Sør-Øst and internal grants from Akershus University Hospital (number: 2679030 and 2699015 to VNK).

Supplementary Information accompanies the paper on British Journal of Cancer website (http://www.nature.com/bjc)

REFERENCES

Aas T, Børresen AL, Geisler S, Smith-Sørensen B, Johnsen H, Varhaug JE, Akslen LA, Lønning PE (1996) Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. *Nat Med* **2**: 811–814

Applanat MP, Buteau-Lozano H, Herve MA, Corpet A (2008) Vascular endothelial growth factor is a target gene for estrogen receptor and contributes to breast cancer progression. *Adv Exp Med Biol* **617**: 437–444

Bardou V-J, Arpino G, Elledge RM, Osborne CK, Clark GM (2003) Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *J Clin Oncol* **21**: 1973–1979

Berns EM, Foekens JA, Vossen R, Look MP, Devilee P, Henzen-Logmans SC, Van Staveren IL, Van Putten WL, Inganas M, Meijer-Van Gelder ME, Cornelisse C, Claassen CJ, Portengen H, Bakker B, Klijn JG (2000) Complete sequencing of TP53 predicts poor response to systemic therapy of advanced breast cancer. *Cancer Res* **60**: 2155–2162

Berns EMJJ, Klijn JGM, Look MP, Grebenchtchikov N, Vossen R, Peters H, Geurts-Moespot A, Portengen H, Van Staveren IL, Meijer-Van Gelder ME, Bakker B, Sweep FCGJ, Foekens JA (2003) Combined vascular endothelial growth factor and TP53 status predicts poor response to tamoxifen therapy in estrogen receptor-positive advanced breast cancer. *Clin Cancer Res* **9**: 1253–1258

Bertheau P, Turpin E, Rickman DS, Espié M, De Reyniès A, Feugeas J-P, Plassa L-F, Soliman H, Varna M, De Roquancourt A, Lehmann-Che J, Beuzard Y, Marty M, Misset J-L, Janin A, De Thé H (2007) Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin – cyclophosphamide regimen. *PLoS Med* **4**: 10

Blandino G, Levine AJ, Oren M (1999) Mutant p53 gain of function: differential effects of different p53 mutants on resistance of cultured cells to chemotherapy. *Oncogene* **18**(2): 477–485

Børresen AL, Andersen TI, Eyfjörd JE, Cornelis RS, Thorlacius S, Borg A, Johansson U, Theillet C, Scherneck S, Hartman S (1995) TP53 mutations and breast cancer prognosis: particularly poor survival rates for cases with mutations in the zinc-binding domains. *Genes Chromosomes Cancer* **14**: 71–75

Buteau-Lozano H, Ancelin M, Lardeux B, Milanini J, Perrot-Applanat M (2002) Transcriptional regulation of vascular endothelial growth factor by estradiol and tamoxifen in breast cancer cells: a complex interplay between estrogen receptors alpha and beta. *Cancer Res* **62**(17): 4977–4984

D'Assoro AB, Leontovich A, Amato A, Ayers-Ringler JR, Quatraro C, Hafner K, Jenkins RB, Libra M, Ingle J, Stivala F, Galanis E, Salisbury JL

(2010) Abrogation of p53 function leads to metastatic transcriptome networks that typify tumor progression in human breast cancer xenografts. *Int J Oncol* **37**: 1167–1176

Dawson MI, Chao WR, Pine P, Jong L, Hobbs PD, Rudd CK, Quick TC, Niles RM, Zhang XK, Lombardo A (1995) Correlation of retinoid binding affinity to retinoic acid receptor alpha with retinoid inhibition of growth of estrogen receptor-positive MCF-7 mammary carcinoma cells. *Cancer Res* **55**(19): 4446–4451

Dhar G, Banerjee S, Dhar K, Tawfik O, Mayo MS, Vanveldhuizen PJ, Banerjee SK (2008) Gain of oncogenic function of p53 mutants induces invasive phenotypes in human breast cancer cells by silencing CCN5/WISP-2. *Cancer Res* **68**(12): 4580–4587

Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **8**: 242

Dittmer J (2003) The Biology of the Ets1 Proto-Oncogene. *Mol Cancer* **2**: 29

Enerly E, Steinfeld I, Kleivi K, Leivonen S-K, Aure MR, Russnes HG, Rønneberg JA, Johnsen H, Navon R, Rødland E, Mäkelä R, Naume B, Perälä M, Kallioniemi O, Kristensen VN, Yakhini Z, Børresen-Dale A-L (2011) miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS One* **6**: 13

Fernandez-Cuesta L, Anaganti S, Hainaut P, Olivier M (2010) p53 status influences response to tamoxifen but not to fulvestrant in breast cancer cell lines. *Int J Cancer* **128**(8): 1813–1821

Gjoerup OV, Wu J, Chandler-Militello D, Williams GL, Zhao J, Schaffhausen B, Jat PS, Roberts TM (2007) Surveillance mechanism linking Bub1 loss to the p53 pathway. *Proc Natl Acad Sci USA* **104**: 8334–8339

Goeman JJ, Van Houwelingen HC, Finos L (2011) Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika* **98**: 381–390

Han QX, Allegretto EA, Shao ZM, Kute TE, Ordonez J, Aisner SC, Rishi AK, Fontana JA (1997) Elevated expression of retinoic acid receptor-alpha (RAR alpha) in estrogen-receptor-positive breast carcinomas as detected by immunohistochemistry. *Am J Surg Pathol Part B* **6**(1): 42–48

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**(1): 44–57

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, Von Mering C (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucl Acids Res* **37**: D412–D416

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**: 118–127

Kandioler-Eckersberger D, Ludwig C, Rudas M, Kappel S, Janschek E, Wenzel C, Schlagbauer-Wadl H, Mittlböck M, Gnant M, Steger G, Jakesz R (2000) TP53 mutation and p53 overexpression for prediction of response to neoadjuvant treatment in breast cancer patients. *Clin Cancer Res* **6**: 50–56

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

Kim E, Deppert W (2006) The versatile interactions of p53 with DNA: when flexibility serves specificity. *Cell Death Differ* **13**(6): 885–889

Langerød A, Zhao H, Borgan Ø, Nesland JM, Bukholm IR, Ikdahl T, Kåresen R, Børresen-Dale A-L, Jeffrey SS (2007) TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res* **9**(3): R30

Li S, Banck M, Mujtaba S, Zhou M-M, Sugrue MM, Walsh MJ (2010) p53-Induced Growth Arrest Is Regulated by the Mitochondrial SirT3 Deacetylase. *PLoS One* **5**: 12

Liang Y, Wu J, Stancel GM, Hyder SM (2005) p53-dependent inhibition of progestin-induced VEGF expression in human breast cancer cells. *J Steroid Biochem Mol Biol* **93**(2-5): 173–182

Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* **8**: 431

Liu W, Konduri SD, Bansal S, Nayak BK, Rajasekaran SA, Karuppayil SM, Rajasekaran AK, Das GM (2006) Estrogen receptor-alpha binds p53 tumor suppressor protein directly and represses its function. *J Biol Chem* **281**(15): 9837–9840

Maity A, Pore N, Lee J, Solomon D, O'Rourke DM (2000) Epidermal growth factor receptor transcriptionally up-regulates vascular endothelial growth factor expression in human glioblastoma cells via a pathway involving phosphatidylinositol 3′-kinase and distinct from that induced by hypoxia. *Cancer Res* **60**(20): 5879–5886

Merico D (2009) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**: e13984

Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA* **102**(38): 13550–13555

Naume B, Zhao X, Synnestvedt M, Borgen E, Russnes HG, Lingjaerde OC, Strømberg M, Wiedswang G, Kvalheim G, Kåresen R, Nesland JM, Børresen-Dale A-L, Sørlie T (2007) Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol Oncol* **1**(2): 160–171

Norberg T, Klaar S, Kärf G, Nordgren H, Holmberg L, Bergh J (2001) Increased p53 mutation frequency during tumor progression – results from a breast cancer cohort. *Cancer Res* **61**: 8317–8321

O'Shaughnessy J (2005) Extending survival with chemotherapy in metastatic breast cancer. *Oncologist* **10**(Suppl 3): 20–29

Olivier M, Langerød A, Carrieri P, Bergh J, Klaar S, Eyfjord J, Theillet C, Rodriguez C, Lidereau R, Bièche I, Varley J, Bignon Y, Uhrhammer N, Winqvist R, Jukkola-Vuorinen A, Niederacher D, Kato S, Ishioka C, Hainaut P, Børresen-Dale A-L (2006) The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. *Clin Cancer Res* **12**: 1157–1167

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J clin oncol* **27**(8): 1160–1167

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. *Vienna Austria R Foundation for Statistical Computing* **1:** ISBN 3-900051-07-0

Ross-Innes CS, Stark R, Holmes KA, Schmidt D, Spyrou C, Russell R, Massie CE, Vowler SL, Eldridge M, Carroll JS (2010) Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. *Genes Dev* **24**: 171–182

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article 3

Therneau T, Lumley T (2009) Survival: Survival Analysis Including Penalised Likelihood. R package Version 2.35-8. *R package version* 2.35-8. http://CRAN.R-project.org/package=survival

Trinh XB, Tjalma WA, Vermeulen PB, Van den Eynden G, Van der Auwera I, Van Laere SJ, Helleman J, Berns EM, Dirix LY, van Dam PA (2009) The VEGF pathway and the AKT/mTOR/p70S6K1 signalling pathway in human epithelial ovarian cancer. *Br J Cancer* **100**(6): 971–978

Wyllie F, Haughton M, Bartek J, Rowson J, Wynford-Thomas D (2003) Mutant p53 can delay growth arrest and loss of CDK2 activity in senescing human fibroblasts without reducing p21(WAF1) expression. *Exp Cell Res* **285**(2): 236–242

Yang Y, Kort EJ, Ebrahimi N, Zhang Z, Teh BT (2010) Dual KS: defining gene sets with tissue set enrichment analysis. *Cancer Inform* **9**: 1–9

Supplementary Figures for Paper II

Suppl Figure 1 **-** Flowchart of analysis.

Suppl Figure 2 - Enrichment map showing the differentially enriched pathways according to the *TP53* mutation status.

Each node, a differentially enriched pathway, is color-coded according to the FDR corrected *globaltest* p-values. Node side represents number of genes in the pathway. Thickness of the edges represents the fraction of overlapping genes between any two adjacent nodes (pathways).

Suppl Figure 3 - Venn diagram displays a gray intersection area representing the validated geneset.

Four genesets were inferred by using globaltest and modified KS test by performing analysis on primary and compiled validation datasets *(Suppl Table 3 and 4)*. Validated gene signature of 112 *(Suppl Table 5)* consensus genes consists of ones that are found to be either significantly associated or upregulated in primary and validation datasets. In the Venn diagram these genes are marked by an intersection of blue and red areas with the number of genes in each of the areas shown in bold italics fonts. Numbers of genes shown in the diagram correspond to the number of unique gene identified as significant in each geneset.

Suppl Figure 4 - Class-specific potential candidate genesets by *TP53* mutation status show differential GO enrichment.

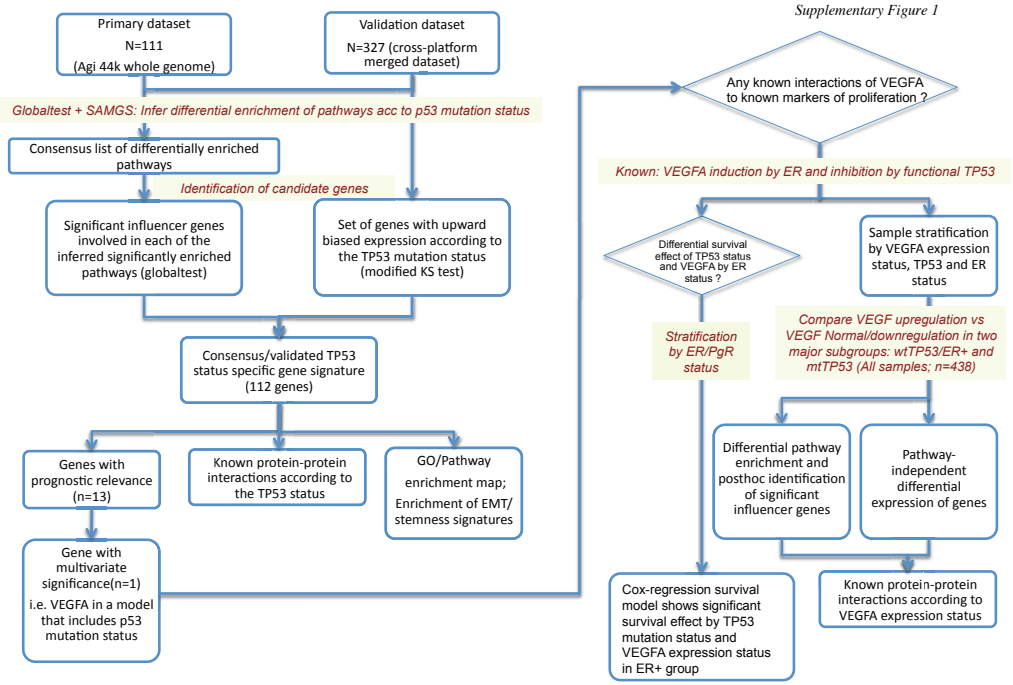
Nodes are color-coded according to the class of their overrepresentation. Nodes in red indicate enriched GO terms in mutant TP53 samples; blue nodes indicate enriched GO terms in wt *TP53* samples. Node size is proportionate of the number of genes assigned to a particular GO term. Thickness of the edges is proportionate to the overlap between the GO terms.

Suppl Figure 5 -Predicted protein-protein functional interaction network corresponding to the differentially expressed genes between wt*TP53* ER+ VEGFA+ samples and wt*TP53* ER+ VEGFA- samples. Besides nodes differentially associated between ER+ VEGFA+ wt *TP53* and ER+ VEGFA- wt*TP53* samples are also shown.

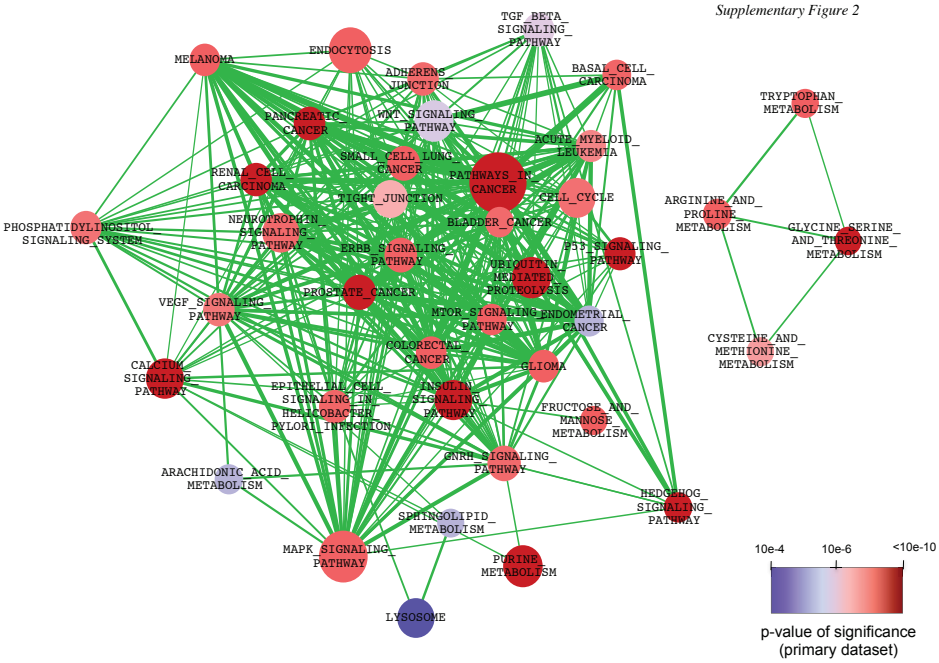
Suppl Figure 6 **-** Proportion of cases with VEGFA upregulation in subgroups based on ER status and *TP53* mutation status. Lowest occurrence of VEGFA upregulation was observed in wtTP53/ER+ samples, but is predictive of poor survival. Thus, in this group of tumors, complex interplay between ER, TP53 and VEGF signaling forms a core biological feature.

Supplementary Figures of Paper II

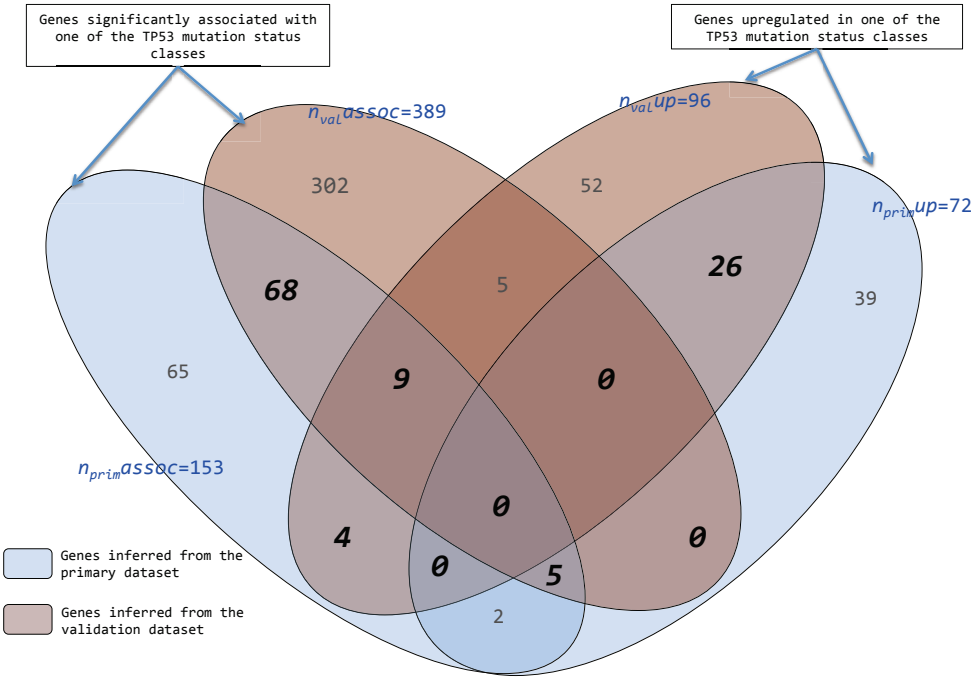
Supplementary Figure 1



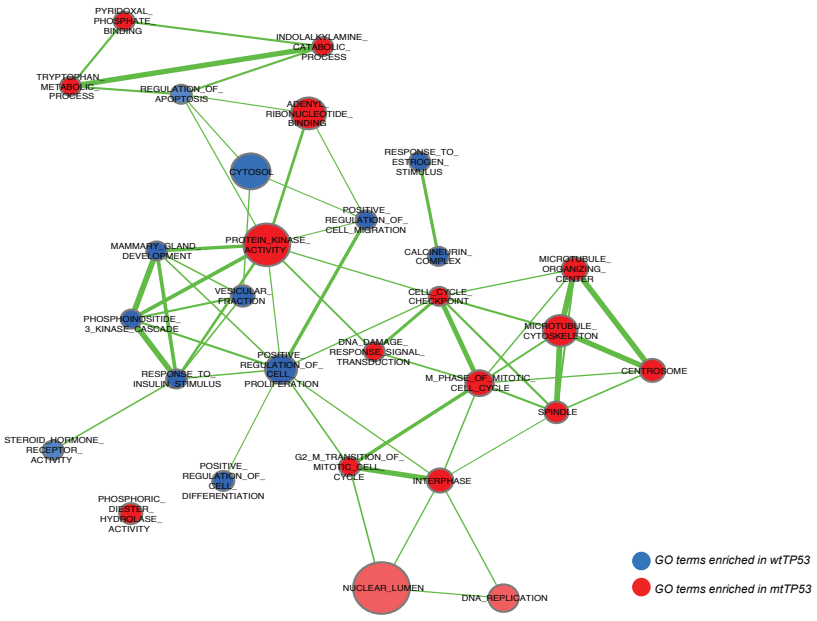
Supplementary Figure 2



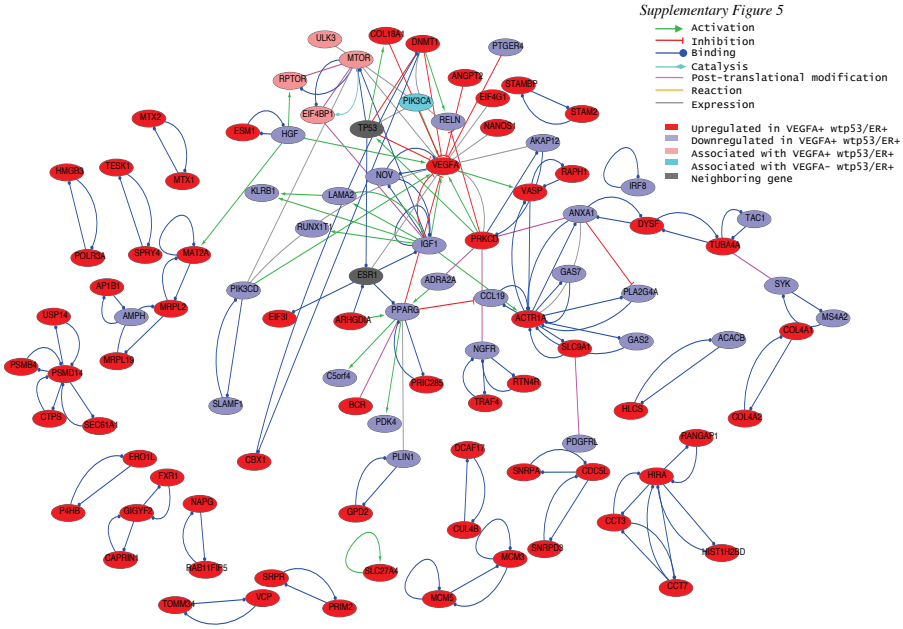
Supplementary Figure 3

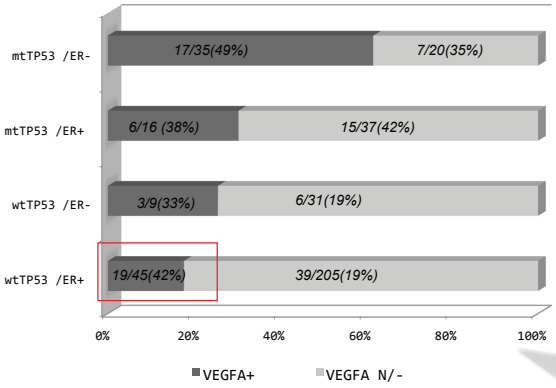


Supplementary Figure 4

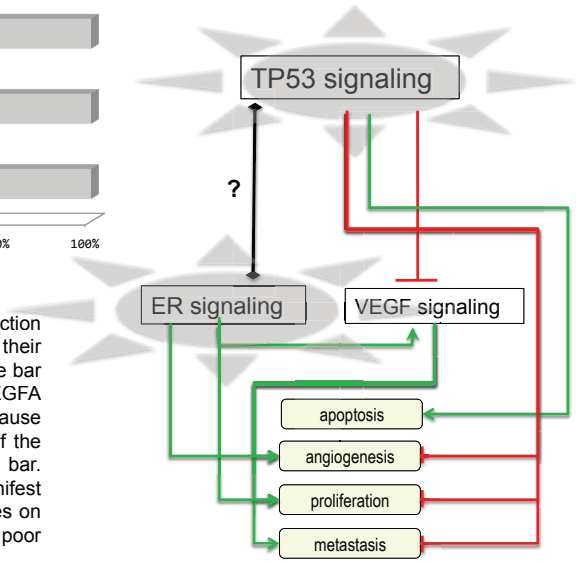


Supplementary Figure 5





Supplementary Figure 6



ER+/wt*TP53* tumors show active interaction between ER and P53 signaling along with their transactivation effects. Dark gray fraction of the bar indicates percentage of samples with VEGFA upregulation. The proportion of patients with cause of death attributed to breast cancer in each of the strata are also mentioned on the respective bar. Subset (here about 18% of all) of tumors manifest with upregulated VEGFA with its consequences on VEGF signaling pathway and ultimately poor patient survival.

Supplementary table 1: **Sample characteristics table showing the number and percentage of patients in each TP53 mutation status class with its estrogen receptor status (Percentage of total number of samples in each TP53 status stratified by ER status are shown in the bracket; Last column shows sum of samples in each TP53 status class and in percentage of total samples categorized in that class shown in bracket)**

	Estrogen receptor status			
	ER -ve	**ER +ve**	**Unknown**	
Primary Dataset*: Naume et al 2007; Enerly et al, 2010*				
wild-type *TP53*	16(38)	55(82)	2(100)	73(66)
mutant *TP53*	26(62)	12(18)	0(0)	38(34)
	42	67	2	*N=111*
Validation Dataset I: *Langerød et al, 2007*				
wild-type TP53	12(44.5)	36(88)	8(100)	56(73.5)
mutant *TP53*	15(55.5)	5(12)	0(0)	20(26.5)
	27	41	8	*N=76*
Validation Dataset II : *Miller et al. 2003*				
wild-type TP53	15(44)	174(81.5)	4(100)	193(77)
mutant *TP53*	19(56)	39(18.5)	0(0)	58(23)
	34	213	4	*N=251*
Combined clinical dataset				
wild-type TP53	43(41.5)	265(82.5)	14(100)	322(73.5)
mutant *TP53*	60(58.5)	56(17.5)	0(0)	116(26.5)
	103	321	14	*N=438*

Supplementary Table 2: Table shows differentially enriched pathways identified by comparing two classes- presence or absense of TP53 mutation in BC expression profiles in primary and validation datasets. The analysis was performed by using two tests- globaltest and SAMGS. The list of significant pathways at Benjamini-Hochberg(BH) adjusted p-value cut-off level <1.0e-4 in primary datasets are shown in the table, sorted by BH adjusted p-value. The signficant pathways based on primary datasets were also found significant on validation dataset.

		Primary dataset				***Validation dataset***			
		Asymptotic global test		***SAM-GS***		***Asymptotic global test***		***SAM-GS***	
KEGGID	**KEGG PATHWAY NAME**	**P-value asymptotic globaltest**	**BH corrected p-value**	**SAMGS p-value**	**FDR adj p-value**	**P-value asymptotic globaltest**	**BH corrected p-value**	**SAMGS p value**	**FDR adj p-value**
hsa:00230	Purine metabolism	1.6e-11	1.8e-09	<10e-6	<10e-6	6.8e-35	2.4e-36	<10e-6	<10e-6
hsa:04115	p53 signaling pathway	2.4e-11	1.8e-09	<10e-6	<10e-6	4.3e-33	2.4e-34	<10e-6	<10e-6
hsa:05211	Renal cell carcinoma	6.5e-11	3.7e-09	<10e-6	<10e-6	4.2e-17	1.3e-17	<10e-6	<10e-6
hsa:05200	Pathways in cancer	3.3e-10	1.1e-08	<10e-6	<10e-6	6.5e-28	6.9e-29	<10e-6	<10e-6
hsa:05215	Prostate cancer	4.3e-10	1.1e-08	<10e-6	<10e-6	1.4e-28	1.3e-29	<10e-6	<10e-6
hsa:04540	Gap junction	2.1e-09	3.9e-08	NS	NS	7.4e-38	1.6e-39	<10e-6	NS
hsa:04020	Calcium signaling pathway	2.3e-09	4.1e-08	<10e-6	<10e-6	3.1e-26	4.8e-27	<10e-6	<10e-6
hsa:00260	Glycine, serine and threonine metabolism	3.1e-09	4.7e-08	<10e-6	<10e-6	7.7e-25	1.4e-25	<10e-6	<10e-6
hsa:05212	Pancreatic cancer	4.2e-09	5.7e-08	<10e-6	<10e-6	6.6e-38	1.2e-39	<10e-6	<10e-6
hsa:04340	Hedgehog signaling pathway	4.7e-09	6.0e-08	<10e-6	<10e-6	2.4e-20	5.8e-21	<10e-6	<10e-6
hsa:05222	Small cell lung cancer	7.3e-09	7.9e-08	<10e-6	<10e-6	5.1e-38	6.8e-40	<10e-6	<10e-6
hsa:04120	Ubiquitin mediated proteolysis	1.2e-08	1.2e-07	<10e-6	<10e-6	3.7e-38	3.3e-40	<10e-6	<10e-6
hsa:04910	Insulin signaling pathway	1.2e-08	1.2e-07	<10e-6	<10e-6	3.5e-26	5.8e-27	<10e-6	<10e-6
hsa:00051	Fructose and mannose metabolism	1.3e-08	1.3e-07	<10e-6	<10e-6	3.4e-29	2.5e-30	<10e-6	<10e-6
hsa:05218	Melanoma	1.5e-08	1.4e-07	<10e-6	<10e-6	5.2e-17	1.7e-17	<10e-6	<10e-6
hsa:04150	mTOR signaling pathway	1.9e-08	1.7e-07	<10e-6	<10e-6	5.6e-25	9.7e-26	<10e-6	<10e-6
hsa:00380	Tryptophan metabolism	2.3e-08	2.0e-07	<10e-6	<10e-6	2.2e-08	1.5e-08	<10e-6	<10e-6
hsa:04144	Endocytosis	2.9e-08	2.4e-07	<10e-6	<10e-6	2.5e-23	5.0e-24	<10e-6	<10e-6
hsa:00330	Arginine and proline metabolism	3.3e-08	2.5e-07	<10e-6	<10e-6	4.5e-18	1.3e-18	<10e-6	<10e-6
hsa:05214	Glioma	3.3e-08	2.5e-07	<10e-6	<10e-6	3.7e-14	1.5e-14	<10e-6	<10e-6
hsa:00030	Pentose phosphate pathway	4.0e-08	2.9e-07	NS	NS	1.7e-28	1.6e-29	<10e-6	NS
hsa:04010	MAPK signaling pathway	4.6e-08	3.1e-07	<10e-6	<10e-6	4.3e-33	2.4e-34	<10e-6	<10e-6
hsa:04012	ErbB signaling pathway	6.2e-08	3.7e-07	<10e-6	<10e-6	8.2e-17	2.7e-17	<10e-6	<10e-6
hsa:05220	Chronic myeloid leukemia	5.9e-08	3.7e-07	<10e-6	<10e-6	2.0e-22	4.3e-23	<10e-6	NS
hsa:04520	Adherens junction	7.1e-08	4.0e-07	<10e-6	<10e-6	2.1e-12	9.8e-13	<10e-6	<10e-6
hsa:05217	Basal cell carcinoma	9.7e-08	4.8e-07	<10e-6	<10e-6	1.2e-10	6.5e-11	<10e-6	<10e-6
hsa:00600	Sphingolipid metabolism	1.0e-07	4.9e-07	<10e-6	<10e-6	1.2e-13	4.7e-14	<10e-6	<10e-6
hsa:05120	Epithelial cell signaling in Helicobacter pylori infection	1.3e-07	5.8e-07	<10e-6	<10e-6	2.8e-11	1.5e-11	<10e-6	<10e-6
hsa:04722	Neurotrophin signaling pathway	1.5e-07	6.7e-07	<10e-6	<10e-6	4.7e-21	1.1e-21	<10e-6	<10e-6
hsa:00240	Pyrimidine metabolism	1.6e-07	6.8e-07	<10e-6	<10e-6	1.3e-30	9.0e-32	<10e-6	NS
hsa:04912	GnRH signaling pathway	1.9e-07	8.2e-07	<10e-6	<10e-6	2.0e-17	6.0e-18	<10e-6	<10e-6
hsa:00310	Lysine degradation	2.0e-07	8.2e-07	<10e-6	<10e-6	9.2e-23	1.9e-23	<10e-6	NS
hsa:05219	Bladder cancer	2.0e-07	8.2e-07	<10e-6	<10e-6	5.0e-17	1.6e-17	<10e-6	<10e-6
hsa:05210	Colorectal cancer	3.0e-07	1.2e-06	<10e-6	<10e-6	7.3e-11	3.9e-11	<10e-6	<10e-6
hsa:04070	Phosphatidylinositol signaling system	3.1e-07	1.2e-06	<10e-6	<10e-6	4.4e-12	2.2e-12	<10e-6	<10e-6
hsa:04110	Cell cycle	3.5e-07	1.3e-06	<10e-6	<10e-6	2.4e-26	3.7e-27	<10e-6	<10e-6
hsa:03018	RNA degradation	3.8e-07	1.3e-06	NS	NS	2.1e-27	2.5e-28	<10e-6	<10e-6
hsa:05223	Non-small cell lung cancer	4.3e-07	1.5e-06	NS	NS	2.8e-13	1.2e-13	NS	NS
hsa:04370	VEGF signaling pathway	4.5e-07	1.5e-06	<10e-6	<10e-6	1.4e-07	1.0e-07	<10e-6	<10e-6
hsa:05221	Acute myeloid leukemia	6.2e-07	2.1e-06	<10e-6	<10e-6	1.3e-11	6.4e-12	<10e-6	<10e-6
hsa:00270	Cysteine and methionine metabolism	1.2e-06	3.6e-06	<10e-6	<10e-6	6.8e-25	1.2e-25	<10e-6	<10e-6
hsa:04530	Tight junction	1.8e-06	5.3e-06	<10e-6	<10e-6	2.2e-17	6.9e-18	<10e-6	<10e-6
hsa:04350	TGF-beta signaling pathway	2.6e-06	7.3e-06	<10e-6	<10e-6	1.5e-13	5.9e-14	<10e-6	<10e-6
hsa:04310	Wnt signaling pathway	3.8e-06	1.0e-05	<10e-6	<10e-6	2.9e-18	8.2e-19	<10e-6	<10e-6
hsa:04210	Apoptosis	5.2e-06	1.4e-05	NS	NS	6.0e-13	2.6e-13	<10e-6	<10e-6
hsa:04630	Jak-STAT signaling pathway	5.3e-06	1.4e-05	NS	NS	8.4e-11	4.6e-11	NS	NS
hsa:00590	Arachidonic acid metabolism	5.7e-06	1.5e-05	<10e-6	<10e-6	2.3e-11	1.2e-11	<10e-6	<10e-6
hsa:05213	Endometrial cancer	7.3e-06	1.8e-05	<10e-6	<10e-6	1.7e-06	1.3e-06	<10e-6	<10e-6
hsa:04060	Cytokine-cytokine receptor interaction	9.8e-06	2.3e-05	NS	NS	2.8e-12	1.3e-12	NS	NS
hsa:04142	Lysosome	2.2e-05	4.9e-05	<10e-6	<10e-6	2.0e-17	6.1e-18	<10e-6	<10e-6
hsa:03008	Ribosome biogenesis in eukaryotes	3.4e-05	7.4e-05	NS	NS	5.6e-25	9.7e-26	NS	NS
hsa:00562	Inositol phosphate metabolism	NS	NS	<10e-6	<10e-6	2.4e-20	5.7e-21	<10e-6	<10e-6
hsa:04260	Cardiac muscle contraction	NS	NS	<10e-6	<10e-6	9.8e-29	8.2e-30	<10e-6	<10e-6
hsa:04270	Vascular smooth muscle contraction	NS	NS	<10e-6	<10e-6	1.7e-24	3.3e-25	<10e-6	<10e-6
hsa:04360	Axon guidance	NS	NS	<10e-6	<10e-6	3.8e-32	2.3e-33	<10e-6	<10e-6
hsa:04610	Complement and coagulation cascades	NS	NS	<10e-6	<10e-6	6.4e-11	3.4e-11	<10e-6	<10e-6
hsa:04810	Regulation of actin cytoskeleton	NS	NS	<10e-6	<10e-6	2.0e-17	6.0e-18	<10e-6	<10e-6
hsa:04916	Melanogenesis	NS	NS	<10e-6	<10e-6	3.1e-12	1.5e-12	<10e-6	<10e-6
hsa:05010	Alzheimer's disease	NS	NS	<10e-6	<10e-6	5.8e-25	1.0e-25	<10e-6	<10e-6
hsa:05014	Amyotrophic lateral sclerosis (ALS)	NS	NS	<10e-6	<10e-6	2.9e-11	1.5e-11	<10e-6	<10e-6
hsa:05110	Vibrio cholerae infection	NS	NS	<10e-6	<10e-6	1.8e-10	1.0e-10	<10e-6	<10e-6

Supplementary Table 3. **Genesets representing TP53 mutation class-specific signatures inferred using the analysis of primary dataset**: Sets of genes that were found either upregulated (by modified Kolmogorov Smirnov test) or significantly associated (by globaltest at inheritance cut-off <0.05) to one of the class. Association means probe feature smaller non-zero regression coefficient found to be associated with one of *TP53* mutation status by using globaltest.

Gene signature specific to wtTP53 BC

A: Genes identified as significantly associated or upregulated in wt TP53 BC class using both methods

ProbeID	*Inheritance*	*p-value*	*Symbol*	*ProbeID*	*Inheritance*	*p-value*	*Symbol*
A_32_P161140	6.4E-04	5.9E-07	UGCG	A_23_P110122	1.6E-04	8.9E-07	CCNG2
A_23_P46819	1.4E-04	1.1E-06	BTRC	A_23_P108143	6.4E-06	2.1E-07	GAMT

B: Genes identified as significantly associated to the wt TP53 BC class using globaltest

ProbeID	*Inheritance*	*p.value*	*Symbol*	*ProbeID*	*Inheritance*	*p.value*	*Symbol*
A_23_P382775	1.1E-05	7.2E-08	BBC3	A_24_P305312	5.3E-03	3.7E-05	BBC3
A_24_P224488	1.1E-05	2.0E-07	MAPT	A_23_P83579	5.5E-03	3.1E-04	ARNT2
A_32_P183765	1.9E-05	1.2E-06	ERBB4	A_23_P349416	6.3E-03	7.9E-05	ERBB3
A_24_P19228	3.1E-05	7.1E-07	GAMT	A_23_P202837	6.4E-03	2.7E-04	CCND1
A_32_P6344	1.0E-04	1.6E-06	MAP2K4	A_23_P83192	6.5E-03	9.6E-05	PHPT1
A_23_P22143	1.3E-04	2.1E-06	PDE6B	A_23_P423853	6.6E-03	8.9E-05	ERBB4
A_23_P156402	1.6E-04	5.8E-07	NME5	A_24_P225679	7.4E-03	8.5E-05	IRS1
A_23_P257111	1.7E-04	1.3E-07	FBP1	A_23_P4161	7.5E-03	5.4E-05	ARSG
A_23_P352266	2.7E-04	6.9E-06	BCL2	A_23_P55107	7.6E-03	4.0E-05	ULK2
A_23_P417282	4.3E-04	1.0E-05	IGF1R	A_24_P53976	7.8E-03	1.3E-04	GLUL
A_24_P945147	4.5E-04	3.2E-07	RABEP1	A_32_P32739	7.9E-03	1.9E-04	NAGS
A_24_P12065	5.4E-04	6.3E-06	CCNG2	A_32_P41574	9.0E-03	7.0E-05	PXN
A_24_P935103	5.6E-04	4.2E-06	ADCY9	A_23_P211835	9.1E-03	1.3E-04	AGTR1
A_23_P207699	8.0E-04	7.2E-06	MAPT	A_23_P385105	9.2E-03	3.6E-04	PLCD4
A_24_P339514	9.4E-04	1.8E-04	CYP2B6	A_23_P353905	9.7E-03	3.4E-05	DUSP16
A_24_P339416	9.7E-04	9.0E-06	ARSG	A_23_P167093	9.8E-03	5.2E-05	IDUA
A_23_P35414	1.0E-03	3.1E-05	PPP1R3C	A_24_P227993	1.0E-02	7.4E-05	UBE2I
A_24_P399174	1.1E-03	6.3E-06	RABEP1	A_32_P98298	1.1E-02	8.9E-05	DUSP8
A_23_P10743	1.3E-03	1.5E-05	PDE6B	A_23_P216325	1.3E-02	8.4E-04	ASAH1
A_23_P502047	1.4E-03	4.5E-05	CHRD	A_23_P500381	1.7E-02	1.9E-04	HTR7
A_23_P201731	1.5E-03	2.2E-05	TRAF5	A_24_P80532	1.7E-02	3.5E-04	CCNG2
A_32_P17182	1.6E-03	6.5E-05	THBS1	A_24_P193011	1.8E-02	1.1E-03	CCND1
A_23_P111531	1.7E-03	6.9E-05	GLI3	A_24_P782308	1.9E-02	1.6E-04	NEDD4L
A_23_P166616	1.9E-03	3.8E-05	AGTR1	A_23_P127367	1.9E-02	1.1E-04	POLD4
A_23_P258018	1.9E-03	1.4E-05	MYL5	A_23_P405794	2.1E-02	3.3E-04	NAGS
A_23_P20392	1.9E-03	3.1E-05	PSD3	A_23_P24433	2.1E-02	2.4E-04	CTSF
A_24_P577694	1.9E-03	5.0E-05	ADCY1	A_23_P406187	2.2E-02	5.4E-04	NAGS
A_24_P18146	2.6E-03	3.7E-05	PSD3	A_23_P92042	2.4E-02	7.3E-04	ITPR1
A_32_P205637	2.6E-03	7.4E-05	PARD6B	A_24_P184031	2.6E-02	4.3E-04	PHPT1
A_24_P108311	3.5E-03	3.8E-05	NEDD4L	A_24_P357266	2.7E-02	5.1E-04	GRPR
A_24_P63380	3.8E-03	1.2E-04	BMPR1B	A_23_P308924	3.2E-02	1.1E-04	DUSP16
A_23_P113111	4.1E-03	1.6E-04	AR	A_23_P146990	3.6E-02	7.5E-04	WWP1
A_24_P322474	4.4E-03	5.6E-05	PDE4A	A_24_P397294	4.2E-02	4.9E-06	LTC4S
A_23_P313389	4.5E-03	2.2E-04	UGCG	A_24_P688133	4.5E-02	8.9E-05	PPM1A
A_23_P152115	5.0E-03	1.2E-05	NME3	A_23_P216167	4.6E-02	1.0E-03	PSD3
A_23_P18559	5.0E-03	1.4E-04	INPP4B	A_23_P99442	4.9E-02	2.0E-03	FLT3

C: Genes identified as upregulated in wt TP53 BC class using modified KS test

ProbeID	*Symbol*	*ProbeID*	*Symbol*	*ProbeID*	*Symbol*	*ProbeID*	*Symbol*
A_32_P5251	RARA	A_23_P48339	IFT88	A_23_P93514	C6orf97	A_24_P316257	FLJ36208
A_32_P45168	MEIS3P1	A_23_P42811	AGR3	A_23_P75056	GATA3	A_24_P211420	SPEF1
A_32_P190303	LONRF2	A_23_P422115	C9orf116	A_23_P62831	FAM176B	A_24_P153840	FGD3
A_32_P16007	POTEB	A_23_P420348	POTED	A_23_P502470	IL6ST	A_23_P95594	NAT1
A_24_P923684	SIRT3	A_23_P416395	STC2	A_23_P50167	SLC39A6	A_23_P372234	CA12
A_24_P586712	TPRG1	A_23_P41634	ANKRA2	A_23_P132378	CELSR1	A_23_P32577	DACH1
A_24_P383478	ESR1	A_23_P41487	TBC1D9	A_23_P27734	NPAS1	A_23_P309739	ESR1
A_24_P368575	SLC4A7	A_23_P40280	SPEF1	A_23_P255701	LRRC48	A_23_P29663	ZMYND10
A_24_P330518	CA12	A_23_P381102	CCDC74B	A_23_P212608	CLSTN2	A_23_P140427	EVL
A_23_P148249	THSD4	A_23_P16648	PCSK4				

Gene signature specific to mtTP53 BC

A: Genes identified as significantly associated or upregulated in mtTP53 BC class using both methods

ProbeID	*Inheritance*	*p-value*	*Symbol*	*ProbeID*	*Inheritance*	*p-value*	*Symbol*
A_24_P179400	3.6E-04	6.0E-06	VEGFA	A_23_P253484	1.7E-05	4.5E-07	AADAT
A_23_P259692	6.1E-07	4.3E-10	PSAT1				

B: Genes identified as significantly associated to the wt TP53 BC class using globaltest

ProbeID	*Inheritance*	*p-value*	*Symbol*	*ProbeID*	*Inheritance*	*p-value*	*Symbol*
A_23_P209200	3.2E-06	9.5E-09	CCNE1	A_24_P180680	4.6E-03	8.8E-05	LAPTM4B
A_23_P210726	1.4E-05	1.4E-07	CDC25B	A_23_P1782	4.9E-03	1.0E-04	CD82
A_23_P100344	2.5E-05	2.3E-07	ORC6L	A_23_P214121	6.0E-03	1.6E-04	AMD1
A_24_P103264	3.7E-05	9.8E-07	UGT8	A_23_P55477	6.1E-03	5.0E-05	ADORA2B
A_24_P397107	4.3E-05	1.0E-07	CDC25A	A_23_P68487	6.6E-03	5.7E-04	BMP7
A_24_P916195	4.9E-05	1.4E-07	GTSE1	A_23_P48637	6.7E-03	1.1E-04	HIF1A
A_23_P215790	8.7E-05	6.9E-07	EGFR	A_24_P37441	6.7E-03	5.5E-05	PDK1
A_23_P22224	1.0E-04	2.1E-06	EIF4EBP1	A_23_P65757	7.2E-03	8.7E-05	CCNB2
A_23_P60974	1.3E-04	3.6E-06	TDO2	A_24_P393449	7.3E-03	1.3E-05	DAPK1
A_23_P56898	1.3E-04	8.4E-06	KYNU	A_24_P48856	7.4E-03	4.8E-04	CBS
A_24_P942589	1.3E-04	3.5E-06	UGT8	A_24_P115762	7.5E-03	2.3E-05	CTSC
A_24_P11506	1.5E-04	8.8E-06	KYNU	A_23_P163481	5.0E-02	6.8E-04	BUB1B
A_23_P132019	1.6E-04	3.7E-07	ITCH	A_24_P313504	7.8E-03	1.4E-04	PLK1
A_23_P168651	2.1E-04	8.8E-07	CDK6	A_23_P420196	9.1E-03	7.2E-05	SOCS1
A_23_P93641	2.3E-04	1.9E-05	AKR1B10	A_23_P7144	9.4E-03	6.7E-04	CXCL1
A_23_P85783	2.3E-04	1.2E-05	PHGDH	A_23_P7636	9.7E-03	1.5E-04	PTTG1
A_24_P166663	2.7E-04	4.4E-06	CDK6	A_24_P414999	9.8E-03	2.0E-04	LAPTM4B
A_23_P58321	2.7E-04	5.7E-07	CCNA2	A_24_P200000	1.0E-02	1.2E-05	STEAP3
A_23_P124417	2.7E-04	9.6E-07	BUB1	A_23_P140256	1.1E-02	8.5E-05	PNP
A_23_P112026	2.8E-04	2.4E-05	IDO1	A_24_P234196	1.1E-02	1.3E-04	RRM2
A_23_P208126	3.6E-04	7.2E-06	SERPINB5	A_23_P205828	1.1E-02	1.7E-05	TJP1
A_23_P121423	3.6E-04	6.7E-07	CDC25A	A_23_P155335	1.1E-02	1.1E-04	PLD1
A_23_P50081	3.7E-04	9.9E-06	IMPA2	A_23_P79398	1.2E-02	1.8E-05	IL1R2
A_24_P212086	3.8E-04	2.5E-06	SERPINB5	A_24_P371962	1.5E-02	3.6E-04	AMD1
A_23_P259586	4.5E-04	3.8E-06	TTK	A_23_P166306	1.5E-02	2.2E-03	CBS
A_23_P57588	4.9E-04	5.4E-06	GTSE1	A_23_P161297	1.5E-02	6.6E-04	OGDHL
A_23_P149200	4.9E-04	7.3E-06	CDC20	A_23_P123478	1.6E-02	7.8E-05	PDE7A
A_23_P116123	5.3E-04	8.2E-06	CHEK1	A_23_P60032	1.6E-02	3.7E-04	E2F1
A_23_P70398	6.4E-04	1.7E-05	VEGFA	A_23_P28953	1.6E-02	6.5E-04	DNMT3B
A_23_P147431	4.1E-03	4.7E-05	LYN	A_24_P410363	1.7E-02	4.6E-05	EGLN1
A_24_P180654	6.7E-04	5.2E-06	CREB3L2	A_23_P343935	1.7E-02	6.6E-06	EGLN1
A_23_P359245	7.9E-04	3.3E-05	MET	A_32_P72447	1.8E-02	2.6E-04	UBE2S
A_24_P129341	9.8E-04	1.0E-04	AKR1B10	A_32_P171328	1.8E-02	2.2E-04	UBE2S
A_23_P81805	1.2E-03	2.9E-05	VEGFA	A_23_P104493	1.8E-02	4.6E-04	PAPSS2
A_23_P170037	1.5E-03	7.2E-06	MID1	A_23_P571	2.2E-02	6.3E-04	SLC2A1
A_23_P119916	1.7E-03	9.8E-05	WNT6	A_23_P370989	2.3E-02	6.7E-06	MCM4
A_23_P65651	2.0E-03	4.6E-05	WARS	A_23_P103720	2.4E-02	1.5E-04	AGMAT
A_23_P53476	2.0E-03	7.9E-05	LDHB	A_23_P393531	2.4E-02	6.4E-05	INPP4A
A_23_P502520	2.1E-03	1.1E-04	IL4I1	A_23_P102117	2.4E-02	4.4E-04	WNT10A
A_23_P124095	2.1E-03	3.7E-04	CALML5	A_23_P46928	2.5E-02	2.0E-05	PFKP
A_23_P102113	2.3E-03	7.3E-05	WNT10A	A_24_P283288	2.5E-02	3.5E-05	MAPK14
A_24_P91566	2.3E-03	1.1E-05	BMP7	A_23_P254522	2.6E-02	2.1E-06	COL4A4
A_23_P57379	2.3E-03	2.9E-05	CDC45L	A_24_P297539	2.7E-02	7.3E-04	UBE2C
A_23_P250914	2.5E-03	9.9E-05	ATP6V1C2	A_23_P74269	2.8E-02	1.7E-04	SRM
A_23_P408955	2.9E-03	8.4E-05	E2F2	A_23_P160618	2.9E-02	1.1E-04	SH2D2A
A_23_P118174	3.0E-03	1.6E-05	PLK1	A_23_P377197	2.9E-02	8.6E-05	MRAS
A_23_P45799	3.1E-03	1.0E-05	ORC1L	A_32_P184933	3.0E-02	4.3E-04	UBE2S
A_23_P209778	3.2E-03	6.6E-06	POLR2D	A_32_P181131	3.5E-02	1.1E-04	AK3L1
A_23_P31921	3.2E-03	1.0E-04	ASS1	A_23_P110725	3.6E-02	2.6E-04	PRKAA1
A_23_P252163	3.4E-03	3.4E-05	DAPK1	A_24_P376556	3.6E-02	6.0E-05	CYCS
A_23_P28898	3.4E-03	7.6E-05	PLCB4	A_23_P80098	3.7E-02	1.4E-04	GART
A_23_P72747	3.6E-03	4.6E-05	UGT8	A_24_P56388	3.8E-02	8.1E-04	HIF1A
A_23_P10614	3.7E-03	1.4E-05	PDK1	A_23_P29330	4.0E-02	1.7E-04	SMC1B
A_23_P122197	4.1E-03	5.6E-05	CCNB1	A_23_P118815	4.0E-02	1.1E-03	BIRC5
A_23_P146456	7.4E-04	2.9E-06	CTSL2	A_23_P127525	4.1E-02	9.8E-05	ETS1
A_23_P106675	4.1E-03	6.3E-05	PLCG2	A_23_P217339	4.6E-02	8.8E-04	PRKX
A_24_P77082	4.2E-03	8.2E-05	KMO	A_23_P29773	7.5E-03	6.6E-06	LAMP3
A_23_P18579	4.4E-03	5.8E-05	PTTG2				

C: Genes identified as upregulated in mt TP53 BC class using modified KS test

ProbeID	*Symbol*	*ProbeID*	*Symbol*	*ProbeID*	*Symbol*	*ProbeID*	*Symbol*
A_32_P84084	MTSS1L	A_24_P335620	SLC7A5	A_23_P71989	UPP1	A_23_P18135	MRPS25
A_32_P77989	NETO2	A_24_P306214	KIAA1609	A_23_P71170	TRPV6	A_23_P168259	ULBP2
A_32_P32391	OR7E156P	A_24_P277576	TRIP13	A_23_P70448	HIST1H1A	A_23_P160537	C1orf135
A_32_P113784	SOX11	A_24_P213924	KIAA1609	A_23_P50990	CENPO	A_23_P157793	CA9
A_24_P93901	SIN3B	A_24_P205604	PADI2	A_23_P415510	LAD1	A_23_P145485	ULBP2
A_24_P873688	CENPN	A_24_P193648	GPT2	A_23_P381945	KRT7	A_23_P112159	EIF2C2
A_24_P722155	LOC100128098	A_24_P187970	PADI2	A_23_P355075	CENPN	A_23_P1043	C1orf106
A_24_P411749	GPR126	A_24_P165450	TTLL7	A_23_P251730	ATP11C	A_23_P88873	GAN
A_24_P384018	OR7E156P	A_23_P92261	ECE2	A_23_P22378	SOX11	A_23_P210581	KCNG1

Supplementary Table 4. Genesets representing TP53 mutation class-specific signatures inferred using the analysis of validation dataset: Sets of genes that were found either upregulated (by modified Kolmogorov Smirnov test) or significantly associated (by globaltest at inheritance cutoff <0.05) to one of the class. Association means probe feature smaller non-zero regression coefficient found to be associated with one of *TP53* mutation status by using globaltest.

Gene signature specific to wtTP53 BC

A: Genes identified as significantly associated or upregulated in wt TP53 BC class using both methods

UniqueID	Inheritance	p-value	Symbol	UniqueID	Inheritance	p-value	Symbol
Hs.101174	6.8E-15	2.2E-16	MAPT	Hs.439726	1.0E-11	2.2E-13	LAMB2
Hs.471508	4.2E-13	7.3E-15	IRS1	Hs.445000	1.6E-05	1.8E-11	PTGER3
Hs.185677	4.2E-13	7.3E-15	NEDD4L	Hs.657729	9.7E-12	1.0E-12	LRP2

B: Genes identified as significantly associated to the wt TP53 BC class using globaltest

UniqueID	Inheritance	p.value	Symbol	UniqueID	Inheritance	p.value	Symbol
Hs.567295	7.1E-12	2.3E-13	ITPR1	Hs.54941	6.9E-03	4.0E-05	PHKA2
Hs.81131	3.0E-10	2.6E-11	GAMT	Hs.150718	7.2E-03	1.6E-04	JAM3
Hs.390729	7.7E-10	3.4E-11	ERBB4	Hs.405961	8.4E-03	2.6E-04	CREB3L1
Hs.476273	8.9E-10	1.5E-11	CACNA2D2	Hs.494312	8.4E-03	1.9E-04	NTRK2
Hs.494496	2.4E-09	3.2E-10	FBP1	Hs.32959	9.5E-03	7.6E-05	GRK4
Hs.597664	1.6E-08	4.3E-10	IKBKB	Hs.505545	1.0E-02	5.7E-05	SLC11A2
Hs.304249	3.1E-08	1.6E-09	UGCG	Hs.370771	1.1E-02	4.6E-04	CDKN1A
Hs.417962	6.3E-08	2.1E-09	DUSP4	Hs.481022	1.1E-02	5.4E-04	SFRP2
Hs.496240	8.1E-08	4.3E-09	AR	Hs.655455	1.1E-02	3.1E-04	PVRL2
Hs.598475	8.9E-08	1.2E-08	BMPR1B	Hs.298654	1.2E-02	2.4E-04	DUSP6
Hs.77810	1.2E-07	1.1E-09	NFATC4	Hs.643802	1.3E-02	2.5E-04	BTRC
Hs.476358	1.8E-07	4.9E-09	CACNA1D	Hs.434375	1.3E-02	7.8E-09	PTPRB
Hs.200841	3.4E-07	1.2E-08	LAMA2	Hs.49774	1.3E-02	2.3E-05	PTPRM
Hs.212088	6.5E-07	2.8E-08	EPHX2	Hs.523852	1.3E-02	1.6E-03	CCND1
Hs.352298	6.8E-07	2.4E-08	PDGFD	Hs.241575	1.4E-02	1.5E-04	GNPTG
Hs.509067	6.8E-07	1.7E-08	PDGFRB	Hs.654400	1.4E-02	2.7E-04	IMPDH2
Hs.592317	7.0E-07	2.2E-08	TGFB3	Hs.567268	1.6E-02	4.6E-04	FGF7
Hs.160562	7.4E-07	5.4E-08	IGF1	Hs.648394	1.6E-02	4.8E-08	EIF4B
Hs.98367	7.5E-07	4.6E-09	SOX17	Hs.207776	1.6E-02	1.8E-04	AGA
Hs.503163	1.0E-06	7.0E-09	PDE2A	Hs.128433	1.6E-02	2.7E-04	HPGDS
Hs.212606	1.4E-06	5.4E-08	GLS2	Hs.471675	1.7E-02	7.7E-04	DGKD
Hs.517227	2.4E-06	3.8E-08	JAM2	Hs.149261	1.9E-02	5.2E-05	RUNX1
Hs.11590	2.8E-06	1.5E-08	CTSF	Hs.591464	1.9E-02	5.0E-04	CGN
Hs.643120	2.9E-06	1.2E-07	IGF1R	Hs.431101	2.0E-02	1.6E-04	GNG12
Hs.664080	3.3E-06	1.6E-08	RASA1	Hs.30213	2.0E-02	6.9E-07	CLN5
Hs.603842	4.3E-06	9.7E-08	MAGI2	Hs.69089	2.0E-02	1.8E-03	GLA
Hs.591336	4.4E-06	8.5E-08	SESN1	Hs.280987	2.0E-02	2.9E-04	MSH3
Hs.477887	4.6E-06	2.1E-07	AGTR1	Hs.9914	2.1E-02	5.7E-05	FST
Hs.89560	6.3E-06	8.2E-08	IDUA	Hs.196384	2.1E-02	1.2E-03	PTGS2
Hs.434255	6.9E-06	2.8E-07	PSD3	Hs.55999	2.2E-02	1.5E-03	NKX3-1
Hs.460109	8.5E-06	2.4E-07	MYH11	Hs.510225	2.2E-02	2.3E-04	RPS6KA5
Hs.1565	9.4E-06	1.0E-07	NEDD4	Hs.482562	2.4E-02	3.7E-04	F2R
Hs.471404	1.0E-05	1.8E-07	STK36	Hs.475896	2.4E-02	1.0E-06	PDCD6IP
Hs.370854	1.1E-05	1.9E-07	TSC1	Hs.593446	2.4E-02	4.9E-04	FRS2
Hs.234898	1.8E-05	6.1E-10	ACACB	Hs.415768	2.4E-02	1.3E-03	NGFR
Hs.514681	1.8E-05	4.6E-07	MAP2K4	Hs.465744	2.7E-02	3.4E-04	INSR
Hs.21509	3.9E-05	1.5E-06	GLI3	Hs.433738	2.9E-02	3.5E-04	GGT7
Hs.192215	4.1E-05	1.3E-06	ADCY1	Hs.9701	2.9E-02	1.5E-03	GADD45G
Hs.523930	4.1E-05	9.2E-07	TRAF5	Hs.2128	2.9E-02	3.9E-04	DUSP5
Hs.482660	4.6E-05	1.5E-07	ZFYVE16	Hs.169378	3.0E-02	1.7E-04	MPDZ
Hs.527412	4.7E-05	1.7E-06	ASAH1	Hs.111867	3.1E-02	5.0E-04	GLI2
Hs.460238	5.8E-05	4.7E-07	SH3GLB2	Hs.500409	3.2E-02	5.5E-04	GLUD1
Hs.65735	5.9E-05	7.4E-07	PHKG2	Hs.232375	3.2E-02	7.4E-04	ACAT1
Hs.475506	6.3E-05	2.5E-09	IQSEC1	Hs.514496	3.3E-02	2.8E-04	EXOC7
Hs.514423	6.5E-05	1.5E-08	CACNG4	Hs.195364	3.3E-02	4.8E-04	MLH1
Hs.183109	6.6E-05	8.0E-06	MAOA	Hs.459070	3.4E-02	5.0E-03	ARNT2
Hs.198241	6.6E-05	5.6E-06	AOC3	Hs.655277	3.5E-02	1.8E-03	RPS6KA2
Hs.211426	9.9E-05	5.6E-06	THBS4	Hs.469820	3.5E-02	7.8E-04	RALB
Hs.369089	1.0E-04	3.0E-06	COL4A5	Hs.89901	3.5E-02	6.1E-06	PDE4A
Hs.485572	1.0E-04	2.7E-06	SOCS2	Hs.265829	3.5E-02	1.2E-03	ITGA3
Hs.350475	1.5E-04	1.4E-09	SUMF1	Hs.129206	3.7E-02	4.9E-05	CSNK1G3
Hs.162129	1.5E-04	1.1E-06	RASGRF2	Hs.145586	3.7E-02	1.2E-03	COL4A6
Hs.391860	1.6E-04	2.7E-06	ADCY9	Hs.72912	3.8E-02	1.3E-03	CYP1A1
Hs.515417	1.8E-04	5.3E-06	EGLN2	Hs.421724	3.8E-02	2.1E-03	CTSG
Hs.156527	1.8E-04	3.0E-06	AXIN2	Hs.183713	4.0E-02	7.4E-04	EDNRA
Hs.445884	1.9E-04	9.1E-06	WNT3	Hs.167700	4.1E-02	2.2E-04	SMAD5
Hs.82002	2.1E-04	2.1E-06	EDNRB	Hs.16695	4.2E-02	9.1E-04	UBA7
Hs.102	2.3E-04	1.7E-07	AMT	Hs.600384	4.3E-02	4.4E-04	HGSNAT
Hs.19121	2.5E-04	7.3E-07	AP2A2	Hs.516306	4.5E-02	2.0E-04	PSD4
Hs.592123	2.7E-04	6.5E-06	SREBF1	Hs.524517	4.5E-02	2.6E-04	CSF3R
Hs.410970	3.1E-04	4.3E-06	MYL5	Hs.175343	5.0E-02	3.3E-05	PIK3C2A
Hs.658169	3.2E-04	2.6E-06	SFRP4	Hs.221472	5.0E-02	9.9E-04	FER
Hs.651939	3.2E-04	3.4E-09	MAGI1				
Hs.700338	3.7E-04	6.2E-06	DDB2				
Hs.525401	3.8E-04	5.2E-06	ADCY6				
Hs.171626	5.0E-04	2.2E-06	SKP1				
Hs.499886	5.2E-04	1.2E-05	ALDH3A2				
Hs.106070	5.9E-04	4.1E-06	CDKN1C				
Hs.591968	6.6E-04	1.2E-05	FZD4				
Hs.150749	7.5E-04	3.1E-05	BCL2				
Hs.436367	7.7E-04	1.9E-05	LAMA3				
Hs.31595	8.4E-04	9.9E-06	CLDN11				
Hs.632702	9.6E-04	1.1E-05	GLI1				
Hs.442378	1.0E-03	1.4E-04	CDO1				
Hs.372924	1.1E-03	4.6E-05	CREB3L4				
Hs.518525	1.3E-03	2.3E-05	GLUL				
Hs.1360	1.4E-03	7.3E-05	CYP2B6				
Hs.441072	1.4E-03	2.1E-05	POLR2L				
Hs.1872	1.5E-03	8.9E-05	PCK1				
Hs.368431	1.5E-03	4.8E-06	RUNX1T1				
Hs.372688	1.5E-03	3.0E-05	RHOBTB2				
Hs.118262	1.6E-03	2.1E-05	CACNA1C				
Hs.130036	1.7E-03	3.1E-06	PPM1A				
Hs.589848	1.8E-03	1.1E-04	PARD6B				
Hs.653654	1.8E-03	5.7E-05	MAP3K1				
Hs.523829	1.9E-03	7.4E-05	POLD4				
Hs.74034	2.1E-03	3.0E-06	CAV1				
Hs.476448	2.1E-03	1.8E-06	FLNB				
Hs.435761	2.7E-03	3.7E-05	PIAS3				
Hs.529862	2.8E-03	1.2E-05	CSNK1A1				
Hs.162646	2.8E-03	3.1E-05	PPARG				

Gene signature specific to mtTP53 BC

A: Genes identified as significantly associated or upregulated in mtTP53 BC class using both methods

UniqueID	Inheritance	p-value	Symbol	UniqueID	Inheritance	p-value	Symbol
Hs.169840	2.5E-23	7.6E-30	TTK	Hs.58974	1.5E-18	1.8E-20	CCNA2
Hs.226390	1.4E-18	5.1E-20	RRM2	Hs.591697	2.9E-15	9.0E-17	MAD2L1
Hs.350966	6.9E-19	1.3E-20	PTTG1	Hs.592049	4.0E-16	8.5E-18	PLK1
Hs.388733	2.8E-13	1.8E-15	PNPT1	Hs.93002	3.51E-20	1.41E-21	UBE2C

B: Genes identified as significantly associated to the wt TP53 BC class using globaltest

UniqueID	Inheritance	p-value	Symbol	UniqueID	Inheritance	p-value	Symbol
Hs.494261	6.3E-19	6.0E-20	PSAT1	Hs.128065	3.4E-04	6.3E-07	CTSC
Hs.159118	5.5E-17	1.3E-18	AMD1	Hs.632282	3.8E-04	3.2E-08	SLC25A5
Hs.533573	6.1E-17	8.0E-19	CDC7	Hs.591360	4.3E-04	9.6E-09	CCDC6
Hs.631709	1.1E-16	4.8E-18	RAD51	Hs.292026	4.7E-04	6.3E-06	EIF4E2
Hs.334562	8.1E-16	2.0E-17	CDK1	Hs.512656	5.6E-04	3.8E-05	PSPH
Hs.24529	3.4E-15	4.8E-17	CHEK1	Hs.709	6.0E-04	5.6E-06	DCK
Hs.555956	1.6E-14	1.4E-16	NUDT5	Hs.201446	7.9E-04	2.4E-08	PERP
Hs.517582	2.6E-13	7.1E-15	MCM5	Hs.654401	9.3E-04	8.3E-06	IMPDH1
Hs.386189	4.2E-13	7.7E-15	GTSE1	Hs.163451	9.4E-04	2.4E-06	PYCR1
Hs.438720	6.2E-13	9.8E-15	MCM7	Hs.469022	1.0E-03	5.4E-06	DGUOK
Hs.194698	6.3E-13	1.2E-14	CCNB2	Hs.34012	1.1E-03	2.9E-05	BRCA2
Hs.374378	2.1E-12	3.6E-14	CKS1B	Hs.198072	1.2E-03	8.3E-06	PDE4B
Hs.23348	2.6E-12	3.3E-14	SKP2	Hs.502773	1.2E-03	8.2E-05	ADI1
Hs.26010	3.7E-12	4.8E-13	PFKP	Hs.441498	1.4E-03	5.1E-08	STAM
Hs.460184	3.9E-12	7.3E-14	MCM4	Hs.2256	1.4E-03	8.6E-05	MMP7
Hs.23960	4.5E-12	1.3E-13	CCNB1	Hs.379466	1.5E-03	8.6E-06	UBE2A
Hs.437705	6.1E-12	8.7E-14	CDC25A	Hs.73527	1.7E-03	1.1E-05	CSNK2B
Hs.153752	7.2E-12	1.7E-13	CDC25B	Hs.183671	1.9E-03	1.0E-04	TDO2
Hs.153479	1.3E-11	2.2E-13	ESPL1	Hs.129683	2.0E-03	2.0E-05	UBE2D1
Hs.202672	1.9E-11	4.4E-13	DNMT1	Hs.163776	2.0E-03	1.9E-05	UBE2J1
Hs.477481	3.9E-11	9.2E-13	MCM2	Hs.593413	2.0E-03	3.1E-05	CXCR4
Hs.209983	7.8E-11	3.6E-14	STMN1	Hs.411695	2.0E-03	1.9E-04	HK3
Hs.74405	8.7E-11	8.4E-13	YWHAQ	Hs.654580	2.2E-03	1.6E-05	PRIM2
Hs.444118	1.3E-10	2.3E-12	MCM6	Hs.521693	2.3E-03	4.1E-05	CCNE2
Hs.49760	3.2E-10	6.3E-12	ORC6	Hs.160786	2.5E-03	1.8E-04	ASS1
Hs.492314	4.4E-10	1.7E-11	LAPTM4B	Hs.642615	2.7E-03	3.3E-04	SHC4
Hs.522819	6.1E-10	7.5E-12	IRAK1	Hs.71040	2.8E-03	1.8E-05	AP1M1
Hs.367992	7.2E-10	3.7E-11	IMPA2	Hs.654952	3.0E-03	3.4E-05	POLR2J
Hs.279413	1.3E-09	1.2E-11	POLD1	Hs.410228	3.0E-03	2.8E-05	ORC3
Hs.405958	3.9E-09	1.3E-10	CDC6	Hs.144496	3.1E-03	2.9E-04	GMDS
Hs.207745	4.4E-09	5.4E-11	RBL1	Hs.181326	3.2E-03	2.1E-04	MTMR2
Hs.527119	4.4E-09	2.7E-11	PDE7A	Hs.472101	3.3E-03	1.5E-04	PLCB4
Hs.235116	5.1E-09	2.6E-11	GRK6	Hs.654377	3.6E-03	7.0E-04	LDHC
Hs.654393	5.8E-09	2.0E-10	E2F1	Hs.416848	3.7E-03	1.6E-04	CTSW
Hs.306791	7.3E-09	6.1E-11	POLD2	Hs.519972	3.8E-03	1.4E-05	HLA-F
Hs.533013	2.1E-08	2.4E-09	CBS	Hs.351475	3.8E-03	1.4E-06	POLR2K
Hs.244723	2.8E-08	1.1E-09	CCNE1	Hs.84	3.9E-03	4.2E-05	IL2RG
Hs.492407	3.4E-08	7.7E-10	YWHAZ	Hs.659263	4.4E-03	1.2E-05	PMM2
Hs.411641	3.5E-08	2.3E-09	EIF4EBP1	Hs.22587	4.5E-03	6.5E-06	SSX2IP
Hs.6906	3.5E-08	2.4E-16	RALA	Hs.88778	4.6E-03	2.5E-04	CBR1
Hs.469649	3.9E-08	9.5E-10	BUB1	Hs.408458	5.0E-03	3.8E-05	WWP2
Hs.470907	4.4E-08	2.1E-10	AK2	Hs.514012	5.1E-03	5.4E-05	MAP2K3
Hs.497599	7.5E-08	2.1E-12	WARS	Hs.231367	5.2E-03	5.0E-05	IL2RA
Hs.518448	8.0E-08	3.4E-11	LAMP3	Hs.523836	5.2E-03	3.3E-04	GSTP1
Hs.709893	1.0E-07	4.1E-10	SOS1	Hs.280604	5.5E-03	3.9E-05	PPP3R1
Hs.291363	1.1E-07	1.3E-09	CHEK2	Hs.488293	5.7E-03	2.4E-04	EGFR
Hs.486502	1.1E-07	2.1E-09	NRAS	Hs.16130	5.8E-03	1.4E-04	UBE2O
Hs.467701	1.2E-07	3.5E-09	ODC1	Hs.435051	5.8E-03	1.3E-04	CDKN2D
Hs.179565	1.6E-07	2.3E-09	MCM3	Hs.515130	5.9E-03	5.2E-05	VANGL1
Hs.19400	1.7E-07	1.7E-09	MAD2L2	Hs.656	6.0E-03	1.8E-04	CDC25C
Hs.194148	2.4E-07	3.2E-09	YES1	Hs.458276	6.3E-03	1.5E-06	NFKBIE
Hs.75514	2.6E-07	2.5E-09	PNP	Hs.382865	6.6E-03	1.2E-04	PLD1
Hs.477693	2.6E-07	1.1E-09	NCK1	Hs.388004	6.8E-03	1.0E-04	AHCY
Hs.81848	3.0E-07	5.9E-09	RAD21	Hs.696032	7.1E-03	1.0E-07	PPARD
Hs.412707	3.7E-07	4.0E-09	HPRT1	Hs.571037	7.6E-03	3.7E-08	ACAT2
Hs.524219	4.8E-07	2.0E-13	TPI1	Hs.591571	8.1E-03	9.2E-05	PPP1CB
Hs.470633	6.6E-07	1.1E-08	PDK1	Hs.1183	8.2E-03	1.3E-04	DUSP2
Hs.597656	1.1E-06	3.1E-09	MSH2	Hs.103527	8.6E-03	2.0E-04	SH2D2A
Hs.591054	1.9E-06	5.2E-09	BID	Hs.478199	8.8E-03	6.2E-05	PRKCI
Hs.597216	2.4E-06	1.0E-07	HIF1A	Hs.474949	8.9E-03	2.1E-04	RBX1
Hs.76244	2.8E-06	8.6E-08	SRM	Hs.475688	9.4E-03	1.9E-04	UBE2E2
Hs.154510	3.0E-06	1.4E-08	CBR3	Hs.181301	1.1E-02	2.4E-04	CTSS
Hs.108112	3.0E-06	1.8E-08	POLE3	Hs.487933	1.2E-02	1.3E-04	NT5C3
Hs.529618	3.2E-06	3.1E-08	TFRC	Hs.128420	1.3E-02	5.6E-05	VPS4A
Hs.531818	3.9E-06	9.4E-09	POLR1A	Hs.425777	1.3E-02	2.9E-04	UBE2L6
Hs.390788	4.1E-06	1.4E-07	PRKX	Hs.155247	1.3E-02	1.6E-03	ALDOC
Hs.144197	4.5E-06	3.7E-07	UGT8	Hs.470126	1.3E-02	9.3E-04	KYNU
Hs.103755	5.1E-06	7.6E-08	RIPK2	Hs.469060	1.4E-02	1.5E-04	POLE4
Hs.380277	5.6E-06	1.6E-08	DAPK1	Hs.492333	1.4E-02	1.1E-04	STK3
Hs.95577	6.9E-06	9.3E-08	CDK4	Hs.436527	1.4E-02	4.4E-05	ANAPC1
Hs.170009	7.1E-06	1.5E-07	TGFA	Hs.79353	1.6E-02	1.1E-04	TFDP1
Hs.65758	1.1E-05	1.8E-07	ITPR3	Hs.201897	1.8E-02	1.9E-04	POLA2
Hs.431367	1.1E-05	1.8E-10	VTA1	Hs.584238	1.8E-02	3.3E-03	GLDC
Hs.470804	1.1E-05	3.5E-07	UBE2E3	Hs.82919	1.9E-02	2.4E-04	CUL2
Hs.485717	1.1E-05	3.6E-10	SMAP1	Hs.856	2.1E-02	2.6E-04	IFNG
Hs.73793	1.4E-05	6.9E-07	VEGFA	Hs.272062	2.1E-02	1.3E-04	PTPRF
Hs.178695	1.5E-05	2.5E-07	MAPK13	Hs.696238	2.2E-02	2.6E-04	BIRC2
Hs.75850	1.6E-05	4.3E-07	WASF1	Hs.50640	2.2E-02	4.0E-04	SOCS1
Hs.515840	1.7E-05	3.3E-09	DNMT3A	Hs.512152	2.5E-02	3.8E-04	HLA-G
Hs.514821	1.8E-05	9.9E-07	CCL5	Hs.518530	2.5E-02	3.1E-04	PAK2
Hs.446149	2.3E-05	2.7E-06	LDHB	Hs.500756	2.8E-02	2.3E-03	GOT1
Hs.518774	2.8E-05	2.4E-07	PAICS	Hs.88556	2.9E-02	6.7E-04	HDAC1
Hs.297413	3.7E-05	1.0E-06	MMP9	Hs.27695	2.9E-02	1.0E-03	MID1
Hs.479214	4.7E-05	1.5E-06	CD38	Hs.402773	3.1E-02	1.4E-04	PTPN7
Hs.180142	4.8E-05	1.3E-05	CALML5	Hs.173724	3.1E-02	1.4E-03	CKB
Hs.523012	5.0E-05	3.5E-06	DDIT4	Hs.432574	3.2E-02	2.3E-04	POLR2H
Hs.162777	6.9E-05	4.5E-07	POLE2	Hs.523774	3.2E-02	2.4E-04	EHD1
Hs.526464	7.1E-05	3.5E-08	PML	Hs.134084	3.2E-02	3.3E-04	M6PR
Hs.631580	7.5E-05	4.0E-07	UBA2	Hs.624	3.3E-02	4.6E-04	IL8
Hs.496487	8.3E-05	8.7E-07	ATF4	Hs.198998	3.3E-02	6.3E-04	CHUK
Hs.523718	9.2E-05	3.4E-06	SFN	Hs.436219	3.3E-02	8.7E-04	ALDH1B1

UnigeneID	Inheritance	p.value	Symbol
Hs.168762	3.3E-03	1.7E-05	ULK2
Hs.75262	3.5E-03	4.5E-12	CTSO
Hs.437058	3.6E-03	6.1E-05	STAT5A
Hs.292524	4.0E-03	4.9E-05	CCNH
Hs.515032	4.4E-03	4.4E-05	MKNK2
Hs.2820	4.5E-03	8.6E-05	OXTR
Hs.421649	4.9E-03	2.1E-04	HTR2B
Hs.650382	5.2E-03	1.4E-06	RAB5C
Hs.11392	5.3E-03	3.8E-04	FIGF
Hs.171695	5.5E-03	4.8E-07	DUSP1
Hs.321709	5.7E-03	5.1E-05	P2RX4
Hs.78183	6.2E-03	3.7E-04	AKR1C3
Hs.118681	6.6E-03	5.1E-05	ERBB3

UnigeneID	Inheritance	p-value	Symbol
Hs.484741	1.0E-04	1.9E-06	GMPR
Hs.28914	1.1E-04	1.2E-06	APRT
Hs.17908	1.2E-04	3.3E-06	ORC1
Hs.119882	1.3E-04	4.4E-06	CDK6
Hs.127799	1.7E-04	6.4E-06	BIRC3
Hs.9731	1.8E-04	1.6E-06	NFKBIB
Hs.507162	1.8E-04	1.7E-06	VPS37B
Hs.82201	1.9E-04	1.8E-06	CSNK2A2
Hs.331420	1.9E-04	1.4E-06	PPAT
Hs.502461	2.1E-04	3.2E-06	DGKZ
Hs.75527	2.5E-04	1.9E-06	ADSL
Hs.147433	2.6E-04	9.2E-06	PCNA
Hs.119591	3.3E-04	1.9E-06	AP2S1
Hs.40499	3.3E-04	1.5E-05	DKK1

UnigeneID	Inheritance	p-value	Symbol
Hs.355927	3.4E-02	2.3E-04	VDAC2
Hs.112432	3.5E-02	1.5E-03	AMH
Hs.655552	3.5E-02	2.7E-04	ASAP1
Hs.55279	3.6E-02	7.5E-03	SERPINB5
Hs.659934	3.6E-02	2.0E-03	SESN3
Hs.395482	3.8E-02	7.6E-04	PTK2
Hs.596514	3.9E-02	1.5E-05	ATP6V0B
Hs.78089	3.9E-02	5.5E-04	ATP6V1F
Hs.221889	4.1E-02	8.0E-04	CSDA
Hs.654604	4.7E-02	1.1E-04	PPP5C
Hs.145442	4.7E-02	1.0E-03	MAP2K1
Hs.404914	4.7E-02	5.0E-04	ADAM17
Hs.473927	4.8E-02	1.0E-03	PDE9A

C: Genes identified as upregulated in wt TP53 BC class using modified KS test

UnigeneID	Symbol	UnigeneID	Symbol	UnigeneID	Symbol	UnigeneID	Symbol
Hs.446680	RAI2	Hs.81934	ACADSB	Hs.100686	AGR3	Hs.129452	DACH1
Hs.403171	EFHC1	Hs.595458	MAST4	Hs.8876	NAGS	Hs.208124	ESR1
Hs.634522	CIRBP	Hs.491148	PCM1	Hs.21380	LONRF2	Hs.480819	TBC1D9
Hs.189780	NOSTRIN	Hs.133062	STK32B	Hs.524134	GATA3	Hs.716456	SIRT3
Hs.513871	CYB5D2	Hs.533738	IFT46	Hs.526735	ZMYND10	Hs.532082	IL6ST
Hs.29190	C1orf64	Hs.35096	ZBTB4	Hs.125867	EVL	Hs.356135	MEIS3P1
Hs.78913	CX3CR1	Hs.523468	SCUBE2	Hs.239154	ANKRA2	Hs.387057	THSD4
Hs.523080	ZCCHC24	Hs.208681	BBS4	Hs.654583	RARA	Hs.579264	LRRC48
Hs.642706	FMO5	Hs.283749	RNASE4	Hs.210995	CA12	Hs.233160	STC2
Hs.444767	KIF13B	Hs.406050	DNALI1	Hs.660044	C6orf97	Hs.591847	NAT1
Hs.356416	CBX7	Hs.110298	ACBD4	Hs.187376	IFT88		
Hs.283749	ANG	Hs.657403	C7orf63	Hs.584784	RABEP1		

C: Genes identified as upregulated in mt TP53 BC class using modified KS test

UnigeneID	Symbol	UnigeneID	Symbol	UnigeneID	Symbol
Hs.1594	CENPA	Hs.179718	MYBL2	Hs.498248	EXO1
Hs.83758	CKS2	Hs.615092	NUSAP1	Hs.270845	KIF23
Hs.444082	EZH2	Hs.62180	ANLN	Hs.184339	MELK
Hs.409065	FEN1	Hs.524571	CDCA8	Hs.3104	KIF14
Hs.239	FOXM1	Hs.14559	CEP55	Hs.574492	IL4I1
Hs.308045	NCAPH	Hs.532968	HJURP	Hs.449415	EIF2C2
Hs.370834	ATAD2	Hs.380857	RCC2	Hs.519035	LAD1
Hs.5199	UBE2T	Hs.250822	AURKA	Hs.518997	C1orf106
Hs.584901	GPSM2	Hs.524216	CDCA3	Hs.488240	UPP1
Hs.632586	CXCL10	Hs.651950	NUF2	Hs.146161	ECE2
Hs.436912	KIFC1	Hs.470654	CDCA7	Hs.514527	BIRC5
Hs.486401	CENPW	Hs.514527	EPR1	Hs.513797	SLC7A5

Suppl Table 5. **Validated genesets, representing TP53 mutation class-specific signatures. Consensus genesets shown here are based on an overlapping genes between TP53 class-specific inferred genesets inferred on primary and validation datasets. Method of inference: 1: globaltest; 2: modified KS test; 3: both. Individual genesets inferrred from primary and validation datasets are shown in *supplementary tables 3 and 4*, respectively.**

Gene signature specific to wtTP53 BC						
	Primary Dataset			Validation Dataset		
Gene symbol	Significant feature	gobaltest inheritance	Method of inference	Significant feature	gobaltest inheritance	Method of inference
ADCY1	A_24_P577694	1.9E-03	1	Hs.192215	4.1E-05	1
ADCY9	A_24_P935103	5.6E-04	1	Hs.391860	1.6E-04	1
AGR3	A_23_P42811		2	Hs.100686		2
AGTR1	A_23_P166616	1.9E-03	1	Hs.477887	4.6E-06	1
ANKRA2	A_23_P41634		2	Hs.239154		2
AR	A_23_P113111	4.1E-03	1	Hs.496240	8.1E-08	1
ARNT2	A_23_P83579	5.5E-03	1	Hs.459070	3.4E-02	1
ASAH1	A_23_P216325	1.3E-02	1	Hs.527412	4.7E-05	1
BCL2	A_23_P352266	2.7E-04	1	Hs.150749	7.5E-04	1
BMPR1B	A_24_P63380	3.8E-03	1	Hs.598475	8.9E-08	1
BTRC	A_23_P46819	1.4E-04	3	Hs.643802	1.3E-02	1
C6orf97	A_23_P93514		2	Hs.660044		2
CA12	A_24_P330518		2	Hs.210995		2
CCND1	A_23_P202837	6.4E-03	1	Hs.523852	1.3E-02	1
CTSF	A_23_P24433	2.1E-02	1	Hs.11590	2.8E-06	1
CYP2B6	A_24_P339514	9.4E-04	1	Hs.1360	1.4E-03	1
DACH1	A_23_P32577		2	Hs.129452		2
ERBB3	A_23_P349416	6.3E-03	1	Hs.118681	6.6E-03	1
ERBB4	A_32_P183765	1.9E-05	1	Hs.390729	7.7E-10	1
ESR1	A_24_P383478		2	Hs.208124		2
EVL	A_23_P140427		2	Hs.125867		2
FBP1	A_23_P257111	1.7E-04	1	Hs.494496	2.4E-09	1
GAMT	A_23_P108143	6.4E-06	3	Hs.81131	3.0E-10	1
GATA3	A_23_P75056		2	Hs.524134		2
GLI3	A_23_P111531	1.7E-03	1	Hs.21509	3.9E-05	1
GLUL	A_24_P53976	7.8E-03	1	Hs.518525	1.3E-03	1
IDUA	A_23_P167093	9.8E-03	1	Hs.89560	6.3E-06	1
IFT88	A_23_P48339		2	Hs.187376		2
IGF1R	A_23_P417282	4.3E-04	1	Hs.643120	2.9E-06	1
IL6ST	A_23_P502470		2	Hs.532082		2
IRS1	A_24_P225679	7.4E-03	1	Hs.471508	4.2E-13	3
ITPR1	A_23_P92042	2.4E-02	1	Hs.567295	7.1E-12	1
LONRF2	A_32_P190303		2	Hs.21380		2
LRRC48	A_23_P255701		2	Hs.579264		2
MAP2K4	A_32_P6344	1.0E-04	1	Hs.514681	1.8E-05	1
MAPT	A_24_P224488	1.1E-05	1	Hs.101174	6.8E-15	3
MEIS3P1	A_32_P45168		2	Hs.356135		2
MYL5	A_23_P258018	1.9E-03	1	Hs.410970	3.1E-04	1
NAGS	A_32_P32739	7.9E-03	1	Hs.8876		2
NAT1	A_23_P95594		2	Hs.591847		2
NEDD4L	A_24_P108311	3.5E-03	1	Hs.185677	4.2E-13	3
PARD6B	A_32_P205637	2.6E-03	1	Hs.589848	1.8E-03	1
PDE4A	A_24_P322474	4.4E-03	1	Hs.89901	3.5E-02	1
POLD4	A_23_P127367	1.9E-02	1	Hs.523829	1.9E-03	1
PPM1A	A_24_P688133	4.5E-02	1	Hs.130036	1.7E-03	1
PSD3	A_23_P20392	1.9E-03	1	Hs.434255	6.9E-06	1
RABEP1	A_24_P945147	4.5E-04	1	Hs.584784		2
RARA	A_32_P5251		2	Hs.654583		2
SIRT3	A_24_P923684		2	Hs.716456		2
STC2	A_23_P416395		2	Hs.233160		2
TBC1D9	A_23_P41487		2	Hs.480819		2
THSD4	A_23_P148249		2	Hs.387057		2
TRAF5	A_23_P201731	1.5E-03	1	Hs.523930	4.1E-05	1
UGCG	A_32_P161140	6.4E-04	3	Hs.304249	3.1E-08	1
ULK2	A_23_P55107	7.6E-03	1	Hs.168762	3.3E-03	1
ZMYND10	A_23_P29663		2	Hs.526735		2

Gene signature specific to mutantTP53 BC						
	Primary Dataset			Validation Dataset		
Gene symbol	Significant feature	gobaltest inheritance	Method of inference	Significant feature	gobaltest inheritance	Method of inference
AMD1	A_23_P214121	6.0E-03	1	Hs.159118	5.5E-17	1
ASS1	A_23_P31921	3.2E-03	1	Hs.160786	2.5E-03	1
BIRC5	A_23_P118815	4.0E-02	1	Hs.514527		2
BUB1	A_23_P124417	2.7E-04	1	Hs.469649	3.9E-08	1
C1orf106	A_23_P1043		2	Hs.518997		2
CALML5	A_23_P124095	2.1E-03	1	Hs.180142	4.8E-05	1
CBS	A_24_P48856	7.4E-03	1	Hs.533013	2.1E-08	1
CCNA2	A_23_P58321	2.7E-04	1	Hs.58974	1.5E-18	3
CCNB1	A_23_P122197	4.1E-03	1	Hs.23960	4.5E-12	1
CCNB2	A_23_P65757	7.2E-03	1	Hs.194698	6.3E-13	1
CCNE1	A_23_P209200	3.2E-06	1	Hs.244723	2.8E-08	1
CDC25A	A_24_P397107	4.3E-05	1	Hs.437705	6.1E-12	1
CDC25B	A_23_P210726	1.4E-05	1	Hs.153752	7.2E-12	1
CDK6	A_23_P168651	2.1E-04	1	Hs.119882	1.3E-04	1
CHEK1	A_23_P116123	5.3E-04	1	Hs.24529	3.4E-15	1
CTSC	A_24_P115762	7.5E-03	1	Hs.128065	3.4E-04	1
DAPK1	A_23_P252163	3.4E-03	1	Hs.380277	5.6E-06	1
E2F1	A_23_P80032	1.6E-02	1	Hs.654393	5.8E-09	1
ECE2	A_23_P92261		2	Hs.146161		2
EGFR	A_23_P215790	8.7E-05	1	Hs.488293	5.7E-03	1
EIF2C2	A_23_P112159		2	Hs.449415		2
EIF4EBP1	A_23_P22224	1.0E-04	1	Hs.411641	3.5E-08	1
GTSE1	A_24_P916195	4.9E-05	1	Hs.386189	4.2E-13	1
HIF1A	A_23_P48637	6.7E-03	1	Hs.597216	2.4E-06	1
IL4I1	A_23_P502520	2.1E-03	1	Hs.574492		2
IMPA2	A_23_P50081	3.7E-04	1	Hs.367992	7.2E-10	1
KYNU	A_23_P56898	1.3E-04	1	Hs.470126	1.3E-02	1
LAD1	A_23_P415510		2	Hs.519035		2
LAMP3	A_23_P29773	7.5E-03	1	Hs.518448	8.0E-08	1
LAPTM4B	A_24_P180680	4.6E-03	1	Hs.492314	4.4E-10	1
LDHB	A_23_P53476	2.0E-03	1	Hs.446149	2.3E-05	1
MCM4	A_23_P370989	2.3E-02	1	Hs.460184	3.9E-12	1
MID1	A_23_P170037	1.5E-03	1	Hs.27695	2.9E-02	1
PDE7A	A_23_P123478	1.6E-02	1	Hs.527119	4.4E-09	1
PDK1	A_23_P10614	3.7E-03	1	Hs.470633	6.6E-07	1
PFKP	A_23_P46928	2.5E-02	1	Hs.26010	3.7E-12	1
PLCB4	A_23_P28898	3.4E-03	1	Hs.472101	3.3E-03	1
PLD1	A_23_P155335	1.1E-02	1	Hs.382865	6.6E-03	1
PLK1	A_23_P118174	3.0E-03	1	Hs.592049	4.0E-16	3
PNP	A_23_P140256	1.1E-02	1	Hs.75514	2.6E-07	1
PRKX	A_23_P217339	4.6E-02	1	Hs.390788	4.1E-06	1
PSAT1	A_23_P259692	6.1E-07	3	Hs.494261	6.3E-19	1
PTTG1	A_23_P7636	9.7E-03	1	Hs.350966	6.9E-19	3
RRM2	A_24_P234196	1.1E-02	1	Hs.226390	1.4E-18	3
SERPINB5	A_23_P208126	3.6E-04	1	Hs.55279	3.6E-02	1
SH2D2A	A_23_P160618	2.9E-02	1	Hs.103527	8.6E-03	1
SLC7A5	A_24_P335620		2	Hs.513797		2
SOCS1	A_23_P420196	9.1E-03	1	Hs.50640	2.2E-02	1
SRM	A_23_P74269	2.8E-02	1	Hs.76244	2.8E-06	1
TDO2	A_23_P80974	1.3E-04	1	Hs.183671	1.9E-03	1
TTK	A_23_P259586	4.5E-04	1	Hs.169840	2.5E-23	3
UBE2C	A_24_P297539	2.7E-02	1	Hs.93002	3.51E-20	3
UGT8	A_24_P103264	3.7E-05	1	Hs.144197	4.5E-06	1
UPP1	A_23_P71989		2	Hs.488240		2
VEGFA	A_24_P179400	3.6E-04	3	Hs.73793	1.4E-05	1
WARS	A_23_P65651	2.0E-03	1	Hs.497599	7.5E-08	1

Supplementary Table 6 A: Sources of previously reported EMT and stemness marker genelists that were used for enrichment analysis *(results shown in Table 2)* of stemness and EMT signatures.

Name of the stemness geneset	*Nr of unique genes*	*Detailed source*	*PubMedID of Source Publication*
EMT markers	281	Union of core EMT signature from Taube et al. 2010 and EMT associated genes from Sarrio et al 2008	20713713; 22102611
ESC targets	380	Genes overexpressed in hESCs by at least five studies	17204602
PRC2 targets	654	Common genes between ChIP-baed Su12 targets, Eed targets and H3K27 targets	16630818
iPSC	340	Signature based on metaanalysis with exclusion of cell cycle and proliferation genes	21149740
p53 targets in ESC	549	p53 target genes in experimental model ES culture were mapped by mouse-human ortholog database	20018659

Supplementary Table 6 B: List of pathways involved in sustainance of stemness properties in breast cancer and their correponding genes. This table is partly a subset of gene signatures (Supplementary Table 3) inferred on primary dataset.

probeID	inheritance	p.value	class of association	symbol	PathwayName
A_23_P46819	0.000136	1.05E-06	wtTP53	BTRC	Hedgehog signaling
A_23_P502470	0.00083972	1.60E-07	wtTP53	IL6ST	JAK-STAT signaling
A_23_P46819	0.00105316	1.05E-06	wtTP53	BTRC	wnt signaling
A_23_P502047	0.00135703	4.48E-05	wtTP53	CHRD	TGF Beta signaling
A_23_P111531	0.0017726	6.91E-05	wtTP53	GLI3	Hedgehog signaling
A_32_P17182	0.0019344	6.48E-05	wtTP53	THBS1	TGF Beta signaling
A_24_P63380	0.00371579	1.21E-04	wtTP53	BMPR1B	TGF Beta signaling
A_23_P144096	0.00529956	7.42E-05	wtTP53	CISH	JAK-STAT signaling
A_23_P202837	0.01435174	2.70E-04	wtTP53	CCND1	JAK-STAT signaling
A_23_P202837	0.02304684	2.70E-04	wtTP53	CCND1	wnt signaling
A_23_P46482	0.03522856	8.22E-04	wtTP53	IL20	JAK-STAT signaling
A_24_P193011	0.03775724	1.11E-03	wtTP53	CCND1	JAK-STAT signaling
A_23_P91850	0.00055479	2.18E-06	mtTP53	IL20RB	JAK-STAT signaling
A_23_P119916	0.00170403	9.85E-05	mtTP53	WNT6	Hedgehog signaling
A_23_P102113	0.00227874	7.34E-05	mtTP53	WNT10A	Hedgehog signaling
A_24_P91566	0.00229898	1.12E-05	mtTP53	BMP7	Hedgehog signaling
A_24_P91566	0.00423626	1.12E-05	mtTP53	BMP7	TGF Beta signaling
A_23_P119916	0.00574177	9.85E-05	mtTP53	WNT6	wnt signaling
A_23_P68487	0.00663903	5.73E-04	mtTP53	BMP7	Hedgehog signaling
A_23_P28898	0.01154005	7.62E-05	mtTP53	PLCB4	wnt signaling
A_23_P68487	0.01218241	5.73E-04	mtTP53	BMP7	TGF Beta signaling
A_23_P420196	0.01291511	7.17E-05	mtTP53	SOCS1	JAK-STAT signaling
A_23_P76078	0.0131133	8.25E-05	mtTP53	IL23A	JAK-STAT signaling
A_23_P102117	0.02410461	4.36E-04	mtTP53	WNT10A	Hedgehog signaling
A_23_P127288	0.02940133	1.35E-04	mtTP53	IL2RA	JAK-STAT signaling
A_24_P59667	0.03234319	4.51E-04	mtTP53	JAK3	JAK-STAT signaling
A_23_P217339	0.04590294	8.77E-04	mtTP53	PRKX	Hedgehog signaling

Supplementary table 7: Univariate and multivariate prognostic value of all 112 genes in validated TP53 status-specific gene signatures based on Cox regression model (cut-off of significance level p<0.05). Out of 112 genes, expression values of 47 genes correlate with survival. Corresponding significance (p-value), log odds and its 95% confidence interval have been shown Multivariate model based on all 47 genes that initially showed univariate significance of survival results in only three significant genes in a multivariate model. Out of three genes, VEGFA maintains significance when tested with TP53 status and predicted subtype. Wald test p-values have been mentioned for the final Cox regression model based on two factors that maintained significance in multivariate model: TP53 mutation status and VEGFA expression status. For performing this analysis, expression values of genes were discretized by following the procedure described in materials and methods.

	Factor	Significance	Log odds	95% Confidence Interval
Univariate analysis	**UBE2C**	1.53E-05	2.25	1.56-3.25
	CCNB2	2.49E-05	2.23	1.53-3.23
	IMPA2	4.46E-05	2.17	1.5-3.15
	CCNA2	1.36E-04	2.07	1.43-3.02
	PLK1	1.81E-04	2.03	1.4-2.93
	TTK	2.04E-04	2.05	1.4-2.99
	EIF2C2	2.08E-04	2.03	1.4-2.95
	CDC25A	2.40E-04	2.01	1.38-2.91
	LAPTM4B	4.50E-04	1.95	1.34-2.83
	PTTG1	4.91E-04	1.93	1.33-2.79
	BIRC5	5.20E-04	1.92	1.33-2.77
	STC2	7.51E-04	0.48	0.31-0.73
	VEGFA	7.95E-04	1.90	1.31-2.76
	LRRC48	1.21E-03	0.47	0.3-0.74
	SLC7A5	1.32E-03	1.84	1.27-2.68
	MEIS3P1	1.53E-03	0.49	0.32-0.76
	PDK1	1.53E-03	1.85	1.26-2.7
	E2F1	1.54E-03	1.82	1.26-2.63
	MAPT	1.73E-03	0.52	0.34-0.78
	CDC25B	1.78E-03	1.84	1.26-2.71
	CCNB1	4.30E-03	1.73	1.19-2.53
	EIF4EBP1	6.82E-03	1.69	1.16-2.47
	LAD1	7.65E-03	1.67	1.14-2.42
	AMD1	7.83E-03	1.71	1.15-2.55
	MCM4	8.64E-03	1.66	1.14-2.41
	BUB1	8.90E-03	1.65	1.13-2.41
	SIRT3	9.44E-03	0.57	0.38-0.87
	ANKRA2	9.57E-03	0.57	0.37-0.87
	EVL	1.02E-02	0.56	0.36-0.87
	NAGS	1.07E-02	0.55	0.35-0.87
	CBS	1.53E-02	1.61	1.1-2.37
	IFT88	1.64E-02	0.60	0.4-0.91
	C1orf106	1.66E-02	1.59	1.09-2.33
	PFKP	1.86E-02	1.58	1.08-2.32
	SRM	1.93E-02	1.59	1.08-2.33
	CHEK1	1.93E-02	1.59	1.08-2.35
	IL6ST	2.09E-02	0.59	0.38-0.92
	PRKX	2.25E-02	1.56	1.07-2.29
	CCNE1	3.34E-02	1.52	1.03-2.24
	ARNT2	3.35E-02	0.63	0.41-0.96
	NAT1	3.68E-02	0.66	0.44-0.97
	GLI3	3.86E-02	0.63	0.4-0.98
	RABEP1	4.18E-02	0.64	0.42-0.98
	MYL5	4.22E-02	0.64	0.42-0.98
	RRM2	4.38E-02	1.47	1.01-2.13
	C6orf97	4.66E-02	0.66	0.43-0.99
	UGCG	4.73E-02	0.65	0.42-0.99
Multivariate analysis (all 47 genes with significant effect on survival by using cox proportionate hazard model)	**IMPA2**	0.013	2.04	1.17-3.57
	RRM2	0.019	0.46	0.24-0.88
	VEGFA	0.029	1.72	1.06-2.80
	Others	NS		
Multivariate analysis((TP53 mutation status + VEGFA +IMPA2+RRM2+ Predicted subtype*)	**Overall model significance (Wald test p-value)**	2.0E-06		
	TP53 mutation status	0.03	1.69	1.06-2.72
	VEGFA	0.03	1.56	1.04-2.36
	predicted subtype	NS		
	IMPA2	NS		
	RRM2	NS		
Final multi(bi-)variate model (TP53 mutation status + VEGFA)	**Overall model significance (Wald test p-value)**	2.3E-06		
	VEGFA	3.1E-02	1.54	1.04-2.29
	TP53 mutation status	1.9E-04	2.12	1.43-3.14

* Expression profiles were categorized into five molecular subtypes by using a published algorithm described in Parker *et al.* J Clin Oncol. 2009 Mar 10;27(8):1160-7.

***Suppl Table 8 A*: Set of differentially expressed genes between VEGFA+ and VEGFA normal/- wt*TP53 ER*+ samples**

symbol	unigene_id	logFC	adj.P.Val
VEGFA	Hs.73793	1.10	1.9E-58
ESM1	Hs.129944	0.73	5.0E-19
COL4A2	Hs.508716	0.52	2.1E-06
TK1	Hs.515122	0.48	1.2E-03
EGLN3	Hs.135507	0.43	1.2E-02
CSPG4	Hs.513044	0.43	4.3E-04
GJC1	Hs.532593	0.40	3.1E-07
KCNN2	Hs.98280	0.39	2.4E-03
ANGPT2	Hs.583870	0.39	1.7E-05
ENPEP	Hs.435765	0.39	2.3E-05
HIST1H2BD	Hs.591797	0.36	4.2E-02
COL4A1	Hs.17441	0.36	1.6E-04
GOT1	Hs.500756	0.35	1.9E-04
PRSS8	Hs.75799	0.35	1.9E-02
NANOS1	Hs.591918	0.35	3.1E-02
ELF3	Hs.67928	0.35	2.3E-02
BDKRB2	Hs.654542	0.34	4.3E-03
NDRG1	Hs.372914	0.33	2.4E-02
MRPL14	Hs.311190	0.33	2.1E-04
CDH13	Hs.654386	0.33	2.9E-02
NDUFA4L2	Hs.75069	0.33	3.4E-04
HMGB3	Hs.19114	0.33	3.9E-02
C6orf129	Hs.284207	0.32	3.6E-02
COL18A1	Hs.517356	0.32	3.2E-03
MCAM	Hs.599039	0.32	1.2E-03
MBOAT2	Hs.467634	0.31	1.8E-02
GPR56	Hs.513633	0.31	2.4E-02
USP14	Hs.464416	0.30	1.3E-03
SLC16A3	Hs.500761	0.30	2.7E-02
MCM5	Hs.517582	0.30	4.3E-02
HSP90AB1	Hs.509736	0.30	2.4E-03
PCDH17	Hs.106511	0.29	1.9E-03
FAM83D	Hs.472716	0.29	3.8E-02
DYSF	Hs.252180	0.29	2.0E-03
TMEM63B	Hs.414473	0.29	8.4E-04
GPD2	Hs.512382	0.27	2.3E-03
ALDOA	Hs.513490	0.27	2.4E-02
KCNK5	Hs.444448	0.27	3.4E-02
TUBA4A	Hs.75318	0.27	3.8E-02
NUP155	Hs.547696	0.27	1.4E-02
B4GALT3	Hs.321231	0.27	4.9E-03
RRP12	Hs.434251	0.27	2.5E-03
GPI	Hs.466471	0.27	3.6E-03
TRAF4	Hs.8375	0.27	6.4E-03
ERO1L	Hs.592304	0.27	1.6E-02
IL17RC	Hs.129959	0.27	4.2E-03
GCAT	Hs.54609	0.26	3.0E-02
SPRY4	Hs.323308	0.26	1.2E-02
EGLN2	Hs.515417	0.26	2.9E-02
EFNA4	Hs.449913	0.26	3.8E-02
TLE1	Hs.197320	0.26	2.2E-02
PRKCD	Hs.155342	0.26	2.4E-02
FLT1	Hs.654360	0.26	3.2E-05
APLN	Hs.303084	0.26	3.6E-02
DHTKD1	Hs.104980	0.26	1.1E-02
PPPDE2	Hs.570455	0.26	1.5E-03
SEC61A1	Hs.518236	0.26	5.2E-03
FN3KRP	Hs.31431	0.26	2.0E-03
SLC9A1	Hs.469116	0.25	4.9E-03
BMP8A	Hs.472497	0.25	1.9E-02
CENPA	Hs.1594	0.25	4.4E-02
STRA13	Hs.37616	0.25	2.4E-02
DDT	Hs.656723	0.25	1.8E-02
SNRPD3	Hs.356549	0.25	6.2E-03

N.B. : Table truncated because of the size.
Complete table is available at : http://www.nature.com/bjc/journal/v107/n10/extref/bjc2012461x3.xls

***Suppl Table 8 B*: Set of differentially expressed genes between VEGFA+ and VEGFA normal/- mt*TP53* samples**

symbol	unigene_id	logFC	adj.P.Val
POC5	Hs.432726	-0.33	6.2E-03
CA9	Hs.63287	0.89	3.0E-03
VEGFA	Hs.73793	0.94	1.2E-21

***Supplementary table 9*: GO Terms Overrepresented in genes found differential expressed between VEGFA+ and VEGFA normal/- wtTP53 samples**

GO Category	GO Term	Fraction of genes	Fisher Exact
biological process	blood vessel morphogenesis	3.2	4.0E-04
biological process	positive regulation of mast cell activation during immune response	0.6	9.3E-05
biological process	cell migration	3.4	2.7E-03
biological process	positive regulation of chemotaxis	1	1.3E-03
biological process	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	1.4	2.6E-03
biological process	anti-apoptosis	2.6	6.7E-03
biological process	regulation of vascular endothelial growth factor receptor signaling pathway	0.6	1.7E-03
cellular component	cytosol	13.6	4.8E-07
cellular component	secretory granule	3	1.9E-04
cellular component	MHC class II protein complex	1.2	1.3E-04
cellular component	proteasome complex	1.6	3.0E-04
cellular component	platelet alpha granule lumen	1.2	9.6E-04
molecular function	MHC class II receptor activity	1	1.3E-04
molecular function	serine-type peptidase activity	2.6	1.3E-03
molecular function	nucleoside transmembrane transporter activity	0.6	1.0E-03
molecular function	ubiquitin thiolesterase activity	1.2	1.7E-02

Suppl Table 10 A: Sets of differentially associated genes of mTOR signaling pathway between VEGFA+ and VEGFA normal/- in wtTP53/ER+ samples. Genes are inferred by applying globaltest

probeID	p-value	class of association	symbol
A_24_P179400	4.62E-08	VEGFA+	VEGFA
A_23_P70398	3.45E-07	VEGFA+	VEGFA
A_23_P81805	1.25E-05	VEGFA+	VEGFA
A_24_P12401	2.58E-04	VEGFA+	VEGFA
A_23_P22224	7.40E-03	VEGFA+	EIF4EBP1
A_24_P237265	4.36E-02	VEGFA+	MAPK1
A_23_P206103	6.48E-02	VEGFA+	ULK3
A_24_P156781	6.64E-02	VEGFA+	PIK3R3
A_23_P34606	7.12E-02	VEGFA+	MTOR
A_23_P384499	8.89E-02	VEGFA+	RPTOR
A_23_P92057	2.11E-02	VEGFA-	PIK3CA
A_24_P398572	5.18E-02	VEGFA-	IGF1
A_24_P304423	6.39E-02	VEGFA-	IGF1

Suppl Table 10 B: Sets of differentially associated genes of mTOR signaling pathway between VEGFA+ and VEGFA normal/- in mutant TP53 samples. Genes are inferred by applying globaltest

probeID	p-value	class of association	symbol
A_23_P70398	1.79E-06	VEGFA+	VEGFA
A_24_P179400	9.26E-06	VEGFA+	VEGFA
A_23_P81805	2.35E-05	VEGFA+	VEGFA
A_24_P12401	6.89E-05	VEGFA+	VEGFA
A_23_P16483	2.99E-02	VEGFA+	STK11
A_23_P42935	3.58E-02	VEGFA+	BRAF
A_32_P15017	5.43E-02	VEGFA+	RICTOR
A_23_P110725	6.14E-02	VEGFA+	PRKAA1
A_23_P92057	7.02E-02	VEGFA+	PIK3CA
A_23_P26444	1.51E-02	VEGFA-	MLST8
A_23_P37910	5.40E-02	VEGFA-	MAPK3
A_24_P830690	6.58E-02	VEGFA-	PDPK1

III