Supporting Information

Kristensen et al. 10.1073/pnas.1108781108

SI Materials and Methods

Patient Material. MicMa patients with breast cancer (n = 101) in this study are part of a cohort of patients treated for localized breast cancer from 1995 to 1998, as previously described (1, 2). Samples from the UPPSALA cohort, collected at the Fresh Tissue Biobank, Department of Pathology, Uppsala University Hospital, were selected from a population-based cohort of 854 women diagnosed between 1986 and 2004 with one of three types of primary breast cancer lesions: (*a*) pure DCIS, (*b*) pure invasive breast cancer 15 mm or less in diameter, or (*c*) mixed lesions (invasive carcinoma with an in situ component). The Mammographic Density and Genetics cohort, including 120 healthy women with no malignant disease but some visible density on mammograms, referred to here as healthy women, was included in this study. Two breast biopsies and three blood samples were collected from each woman. The Chin validation set consisted of 113 tumor samples with both expression (GEO accession no. GSE6757) and CGH data (MIAMEExpress accession E-Ucon-1). The UNC validation dataset consisted of 78 tumor samples with both expression (44 K; Agilent Technologies) and SNP-CGH (109 K; Illumina).

MicMa. The 101 patients with breast cancer in this study are part of a cohort previously described (3). Patients treated for localized breast cancer were included in this project (from 1995 to 1998) and have previously been described (4). The routine selection of patients to adjuvant treatment was based on the prevailing national guidelines, where postmenopausal hormone receptor (HR)-positive patients received tamoxifen only, postmenopausal HR-negative patients received 5-fluorouracil (CMF), and premenopausal patients received CMF followed by tamoxifen if HR-positive. Five patients received high-dose chemotherapy, and another 5 received preoperative chemotherapy because of large tumor size. After primary therapy was completed, the patients were followed at 60- to 12-mo intervals. A total of 920 patients were included in the study; clinical correlation to disseminated tumor cells (DTC) status was originally reported for 817 patients (4) and now includes an updated follow-up of 811 patients (median follow-up of 85 mo). Fresh-frozen tissue samples were available from 123 individuals. The study was approved by the Norwegian Regional Committee for Medical Research Ethics, Health Region II (reference no. S-97103). All patients have given written consent for the use of material for research purposes.

The members of the UPPSALA cohort, collected at the Fresh Tissue Biobank, Department of Pathology, Uppsala University Hospital, were selected from a population-based cohort of 854 women diagnosed between 1986 and 2004 with one of three types of primary breast cancer lesions: (*a*) pure DCIS, (*b*) pure invasive breast cancer 15 mm or less in diameter, or (*c*) mixed lesions (invasive carcinoma with an in situ component). All histopathological specimens, both paraffin-embedded (used in immunohistochemical analyses) and frozen (used in microarray and RT-PCR analyses), were reevaluated by a breast pathologist. Thinner sections (4 μm) from the frozen specimens were cut before, between sections 5 and 6, and after the last 20-μm RNA section for H&E staining. These sections were used to estimate the proportion of tumor cells (in situ/invasive cells) in each lesion. Seventy-seven percent of the pure DCIS samples had a DCIS component of >70%. Seventy-six percent of the invasive samples had a tumor content of >70%. Seventy-nine percent of the mixed samples had a tumor/DCIS component of >70%. Invasive breast cancer was classified based on the Elston–Ellis classification system (grade I–III). DCIS lesions were classified according to the European Organization for Research and Treatment of Cancer (EORTC) system. We denoted the EORTC grades I–III using the nomenclature A–C to emphasize that in situ and invasive lesions were classified based on different systems. In lesions with both an invasive element and an in situ element, classification was determined for both elements separately. Control samples of normal breast epithelium were taken from 6 women undergoing surgery for benign conditions, including hyperplasia and fibroadenoma. A total of 109 patient tissues were successfully analyzed by microarrays, of which 31 were pure DCIS, 36 were pure invasive cancers, and 42 were cases of mixed diagnosis. The study was designed to investigate gene expression changes between the groups aiming to identify differences related to tumor progression from in situ to invasive cancer. Patient characteristics are described in Dataset S4. Of the 109 tumors, 29 were removed by mastectomy (12 DCIS, 2 invasive cancers, and 15 mixed cancers) and 80 were removed by breast-conserving surgery (19 DCIS, 34 invasive cancers, and 27 mixed cancers). This study was approved by the Ethics Committee at Uppsala University Hospital.

Mammographic Density and Genetics. The women included in this study had all attended one of six breast diagnostic centers in Norway that are part of the governmentally funded National Breast Cancer Screening Program between 2002 and 2007. Women were eligible if they did not currently use anticoagulants, did not have breast implants, and were not currently pregnant or lactating. A total of 120 healthy women with no malignant disease but some visible density in their mammograms, referred to here as healthy women, were included in this study. Of these, quality-tested expression data were obtained from biopsies from 79 healthy women and array-CGH data (244 K; Agilent Technologies) were available for 81. The women provided information about height, weight, parity, hormone therapy use, and family history of breast cancer. Two breast biopsies and three blood samples were collected from each woman. All women provided signed informed consent. The study was approved by the local ethical committee and local authorities (Institutional Review Board approval no. S-02036).

MicroRNA profiling from total RNA was performed using an Agilent Technologies Human miRNA Microarray Kit (V2) according to the manufacturer's protocol. Scanning on an Agilent Technologies Scanner G2565A and Feature Extraction (FE) v9.5 were used to extract signals. Experiments were performed using duplicate hybridizations (99 samples) on different arrays and time points. Two samples were profiled only once. microRNA signal intensities for replicate probes were averaged across the platform, $\log_2$-transformed, and normalized to the 75th percentile. MicroRNA expression status was scored as present or absent for each gene in each sample by default settings in FE v9.5.

DNA methylation. One microgram of DNA was treated with bisulphite using an EpiTect 96 Bisulfite Kit (Qiagen GmbH). Five hundred nanograms of bisulphite-treated DNA was analyzed using the GoldenGate Methylation Cancer Panel I (Illumina) that simultaneously analyses 1,505 CpG sites in 807 cancer-related genes. At least 2 CpG sites were analyzed per gene, where 1 CpG site is in the promoter region and 1 CpG site is in the first exon. Bead studio software was used for the initial processing of the methylation data according to the manufacturer's protocol. The detection P value for each CpG site was used to validate sample performance, and the dataset was filtered based on the detection

P value, where CpG sites with a detection *P* value >0.05 were omitted from further analysis.

Data preprocessing and PARADIGM parameters. Copy number was segmented using circular binary segmentation (CBS) and then mapped to gene-level measurements by taking the median of all segments that span a RefSeq gene's coordinates in hg18. For mRNA expression, measurements were first probe-normalized by subtracting the median expression value for each probe. The manufacturer's genomic location for each probe was converted from hg17 to hg18 using University of California, Santa Cruz liftOver tool. Per-gene measurements were then obtained by taking the median value of all probes overlapping a RefSeq gene. Methylation probes were matched to genes using the manufacturer's description. PARADIGM was run as it was run previously (5), by quantile-transforming each dataset separately, but data were discretized into bins of equal size rather than at the 5% and 95% quantiles. Pathway files were from the Pathway Interaction Database (6) as previously parsed. Fig. 4 shows summaries of discretized input data, and not integrated pathway level (IPL) values, by counting the fraction of observations in either an up or down bin in each data type and then labeling each node within the bin with the highest fraction of observations in any data type.

HOPACH unsupervised clustering. Clusters were derived using the HOPACH R implementation version 2.10 (7) running on R version 2.12. The correlation distance metric was used with all data types, except for PARADIGM IPLs, which used cosangle because of the nonnormal distribution and prevalence of zero values. For any cluster of samples that contained fewer than five samples, each sample was mapped to the same cluster as the most similar sample in a larger cluster. PARADIGM clusters in the MicMa dataset were mapped to other data types by determining each cluster's mediod (using the median function) in the MicMa dataset and then assigning each sample in another dataset to whichever cluster mediod was closest by cosangle distance. The copy number was clustered on gene-level values rather than by probe. The values that went into the clustering are from the CBS segmentation of each sample. A single value was then generated for each gene by taking the median of all segments that overlap the gene. The samples were then clustered using these gene-level copy number estimates with an uncentered correlation metric in HOPACH. For display, the genes and samples were median-centered.

Kaplain–Meier cluster enrichments. Kaplan–Meier statistics, plots, and cluster enrichments were determined using R version 2.12. Cox *P* values were determined using the Wald test from the coxph() proportional hazards model and log-rank *P* values from a χ^2 test from the survdiff() function. Overall enrichment of a gene's or pathway member's values for a clustering were determined by ANOVA, and enrichment of a gene for a particular cluster label was determined by a *t* test of a gene's values in a particular cluster vs. the gene's values in all other clusters. The false discovery rate was determined using the Benjamini–Hochberg method of p.adjust. The χ^2 values in Dataset S5 were determined using summary.table().

1. Naume B, et al. (2007) Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol Oncol* 1: 160–171.
2. Wiedswang G, et al. (2003) Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer. *J Clin Oncol* 21:3469–3478.
3. Wiedswang G, et al. (2003) Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer. *J Clin Oncol* 21: 3469–3478.
4. Naume B, et al. (2007) Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol Oncol* 1:160–171.
5. Vaske CJ, et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26:i237–i245.
6. Schaefer CF, et al. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res* 37: D674–D679.
7. van der Laan MJ, Pollard KS (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J Stat Planning Inference* 117:275–303.

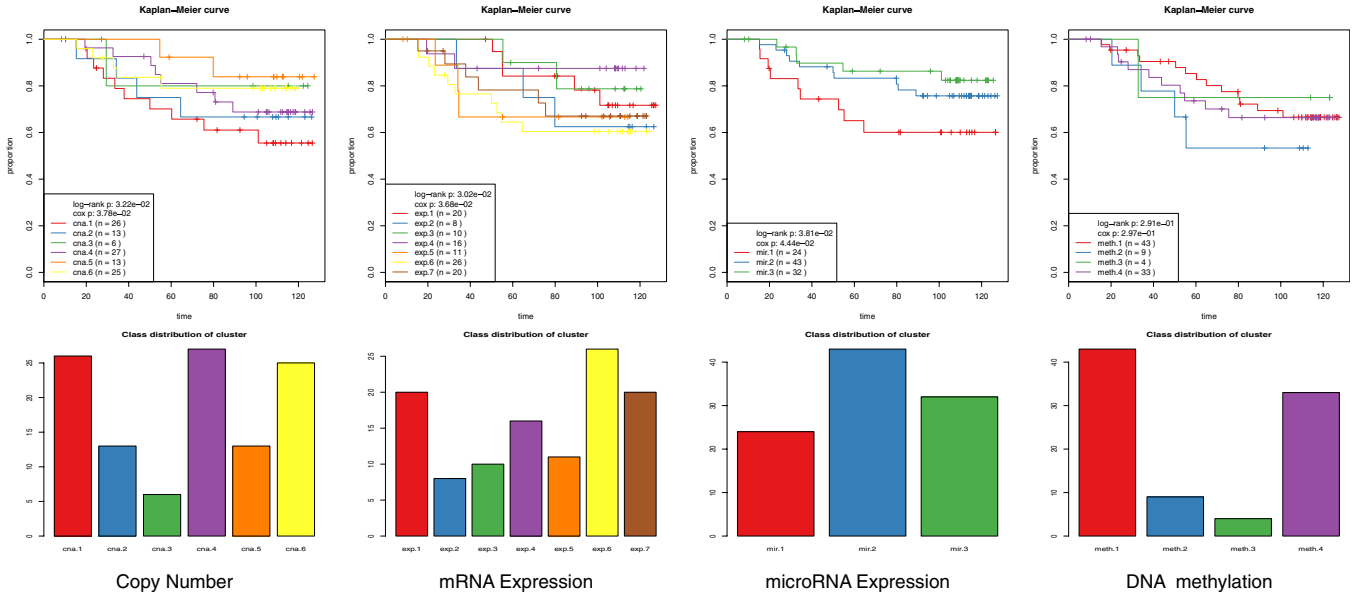

Fig. S1. HOPACH 2.10 clustering of each molecular level (CNA, mRNA, microRNA, and DNA methylation) separately (preintegration). Distribution of unsupervised clusters (HOPACH) and survival curves of the patients of the MicMa cohort according to CNA, mRNA expression, DNA methylation, and microRNA expression. For each type of genomic level, the size of each cluster is plotted on the left, and survival curves are shown on the right.

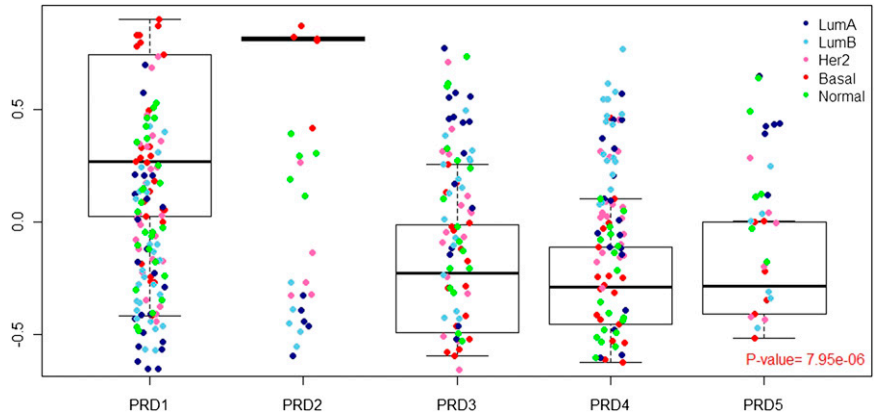

Fig. S2. PDGM clusters and their correlation to the centroids [basal, luminal A (LumA), luminal B (LumB), Her2, and normal-like].

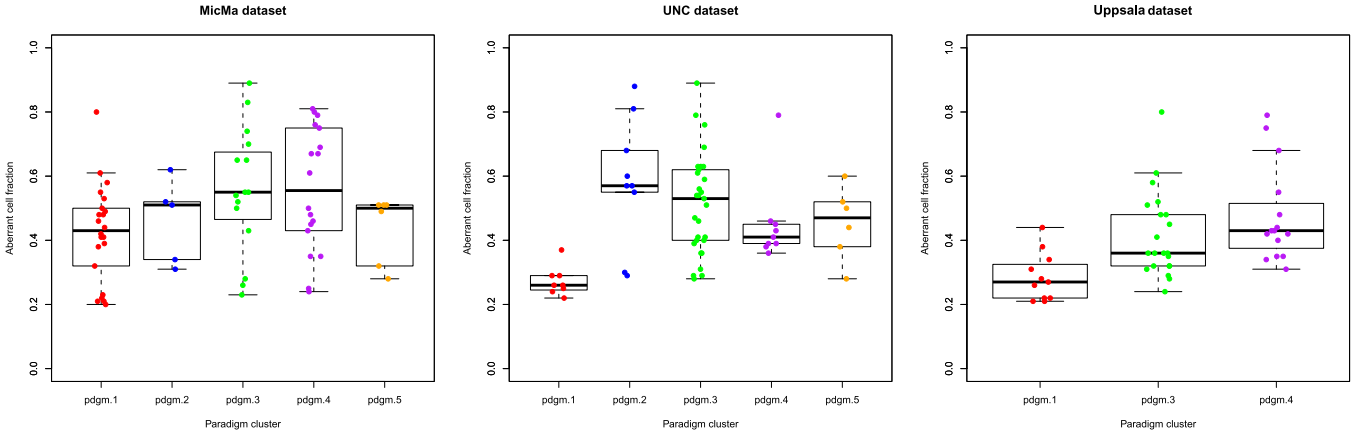

Fig. S3. Difference in tumor content by ASCAT in the different PDGM clusters (box plots). FW.

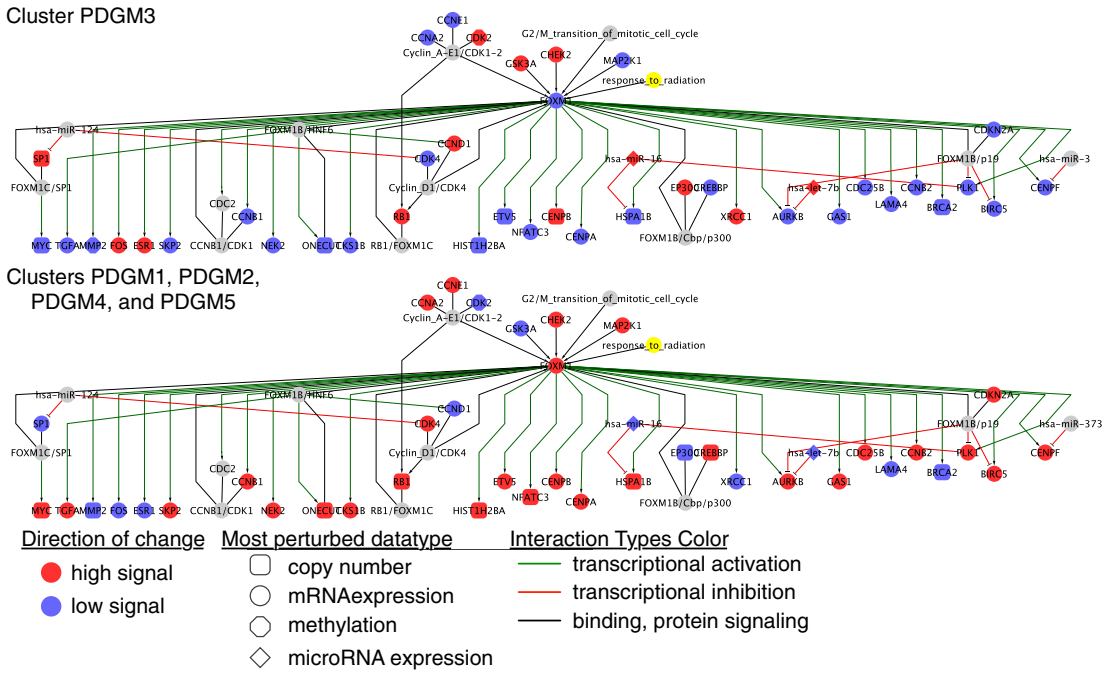

Fig. S4. FOXM1 transcription factor network. (*Upper*) Network diagram summarizes the data from cluster PDGM3. (*Lower*) Cluster summarizes the data from other clusters. Node shapes denote the data type that was most frequently perturbed within each cluster, and node colors denote the direction of perturbation. Edge arrows denote the sign of interactions, and colors denote the type of interaction.

Dataset S1. Most deregulated pathways characterizing the PARADIGM classification in the discovery and validation datasets classified based on their biological function

Dataset S1 (XLSX)

Dataset S2. Breakdown of genes/pathways in the immune-rich PDGM1 based on their biological function

Dataset S2 (XLSX)

Dataset S3. Breakdown of all PDGM-defining genes based on their biological function

Dataset S3 (XLSX)

Dataset S4. Pathways differentially regulated in low/high mammographic breast density

Dataset S4 (XLS)

Dataset S5. Association of the PDGM clusters by clinical and molecular parameters in breast cancer

Dataset S5 (XLSX)

Dataset S6. Difference in tumor content by ASCAT in the different PDGM clusters (*P* values)

Dataset S6 (XLS)