was used per 30 μl PCR. RT-PCR was performed by using the Applied Biosystems AmpliTaq Gold, Cheshire, UK, with either 25 or 30 cycles, each consisting of 30 s at 94°C, 30 s at 55°C, and 45 s at 72°C. PCR products were visualised on 2% Invitrogen agarose E-Gels 96 Gels (Invitrogen Life Technologies, Carlsbad, CA, USA).

Immunohistochemistry and tissue microarray analysis

A cohort of invasive breast carcinomas from 245 patients treated with surgery (wide local excision or mastectomy) and adjuvant anthracycline-based chemotherapy was retrieved from the Department of Histopathology files of the Royal Marsden Hospital (London, UK) with appropriate local Ethical Committee approval. Representative blocks were reviewed by a pathologist (JSRF) and selected cores were incorporated in two duplicate tissue microarray (TMA) blocks [32,33]. Full details of the TMA are given as Additional file 3. TMA samples were dewaxed in xylene, cleared in absolute ethanol and blocked in methanol for 10 minutes. Antigen retrieval for cartilage oligomeric matrix protein (COMP) and IL8 was by boiling slides in citrate buffer (pH 6) for 2 minutes in a pressure cooker, after which they were blocked with normal horse serum (2.5% for 20 minutes; Vector Laboratories VL, Burlingame, CA, USA) and endogenous biotin blocked by pre-incubating with avidin (15 minutes) and biotin (15 minutes). They were then incubated with anti-COMP antibody (1/50; Serotec, Oxford, UK) or IL8 antibody (1/5; Serotec) for 1 h at room temperature. For immunohistochemistry of Periostin (POSTN), sections were pretreated by microwaving in Dakocytomation (Glostrup, Denmark) pH 6 antigen retrieval buffer for 18 minutes, blocked, and anti-POSTN antibody (1/1500; Biovendor Laboratory, Heidelberg, Germany) applied for 30 minutes at room temperature. Antibody binding was detected using Vectastain Universal ABC (VL), visualised with 3,3'-diaminobenzidine DAKO (Corporation, Glostrup, Denmark). Full details on the distribution of ER, PR, HER2, EGFR, CK 14, CK 5/6, and CK 17, as well as P53 (DO7, 1/200; DAKO Corporation) are described elsewhere [33] and summarised in Additional file 3. To evaluate the proliferative activity of tumour cells, immunohistochemical detection of MIB1 antibody to detect Ki-67 nuclear antigen (1/300; DAKO Corporation), which is associated with cell proliferation, was carried out under the same conditions [33]. For these markers, only nuclear staining was considered specific. Ki67 (MIB1) staining was scored low if less than 10% of neoplastic cells were positive, intermediate if 10% to 30% of neoplastic cells were positive and high if more than 30% of neoplastic cells were positive [32]. Tumours were scored positive for P53 if >10% of the nuclei of neoplastic cells displayed strong staining [32].

Cumulative survival probabilities were calculated using the Kaplan-Meier method/log-rank test. Differences between disease-free interval and survival were tested with the log-rank test (two-tailed, confidence interval 95%) using the statistical software Statview 5.0., NC, USA. Multivariate analysis was performed using the Cox multiple hazards model. A P value $<$ 0.05 in the univariate survival analysis was used as the limit for inclusion in the multivariate model.

Results

MPSS analysis of normal luminal and malignant breast cancer cells

The gene expression profiles that were obtained by MPSS analysis yielded 24,288 and 28,404 signature sequences for the malignant and the normal breast epithelium, respectively; these were pared down to the 'signature-centric' version containing 14,245 uniquely mapped and expressed transcripts for the malignant sample and 10,249 transcripts for the normal luminal epithelial sample (Table 1). Based on our HTR (described in Materials and methods [21]), these transcripts corresponded to 8,421 and 6,477 HTR clusters in the malignant and the normal RNA samples, respectively (Table 1), of which 3,191 genes were uniquely expressed in the malignant sample, and 1,297 in the normal sample. To define differential expression, a comparative Poisson test was applied [34] and 6,553 genes were identified that showed a differential expression measurement with $P \leq 0.05$. (Raw and annotated MPSS data are provided as Additional file 4) Expression levels of differentially upregulated transcripts in the tumour sample ranged from less than 10 tpm (*ESR1*, *EGF*, *GPR150*, *GADD45BGIP1*), to over 1,000 tpm (*COL1A1*, *SCGB2A2*, *SELE*, *IL8*).

Establishing a microarray validated transcriptome

The MPSS derived transcriptomes were compared with gene expression profiles of the same RNA pools obtained using three oligonucleotide genome-wide microarrays, Affymetrix U133 Plus 2.0 GeneChip and CodeLink™ Human Whole Genome Bioarray, Agilent Whole Human Genome Oligo Microarray 44 k cDNA array and 20 k brk cDNA microarray. These different microarray platforms were chosen to achieve the highest possible coverage of known transcribed sequences. Features from all platforms were mapped to HTR clusters and our analysis was restricted to those that mapped unambiguously to one HTR cluster. For the Affymetrix platform 41,322 of 54,613 (75.66%) features could be assigned to unique HTR clusters; for CodeLink™ 28,949 of 54,841 (52.78%); for Agilent 32,402 of 44,290 (73.15%); and for the 20 k brk 12,055 of 19,959 (60.4%). Overlay of the transcript coverage of each microarray demonstrated that each platform contributed a set of unique genes as well as those common to other platforms, justifying the use of more than one microarray platform. (Full annotation of each microarray platform to HTR clusters is available as Additional file 5) The microarray features of all four platforms provided a total coverage of 26,103 HTRs, and 6,342 out of 6,553 (96.8%) of the differentially expressed transcripts obtained by MPSS were represented on one or more of these genome-wide platforms.