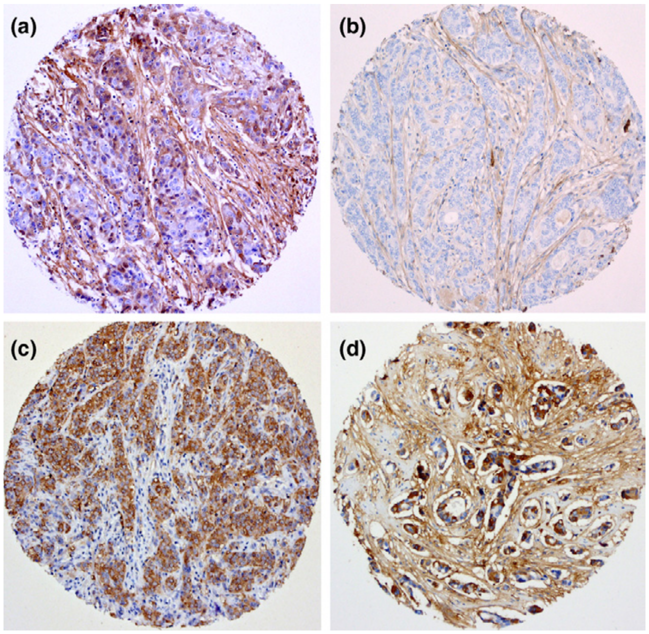
differentially expressed between luminal and myoepithelial cells, and IL8, an inducer of bone resorption. Similarly to POSTN, COMP and IL8 could be clearly detected in the epithelial cells of 21% and 13.9% invasive breast carcinomas, respectively (Figure 5c,d). In contrast to POSTN, however, there was no correlation of COMP or IL8 tumour staining with age, grade, stage, ER, PR, disease-free interval or overall survival, although epithelial expression of the mesenchymal markers POSTN and COMP correlated significantly with each other (Additional file 10).

Discussion

Using highly enriched populations of malignant breast epithelial cells and normal epithelial cells, obtained from immunomagnetic cell sorting, we have established genome-wide molecular signatures specific to the epithelial compartments of both the normal and the malignant human breast. Combining gene profiles obtained from different expression platforms, including direct high-throughput sequencing (MPSS) and multiple microarray platforms, yielded a validated transcriptome comprising 8,051 differential transcripts. These data provide a basis for the molecular changes that occur in the transition from normal luminal to malignant epithelial cells, and also allow further analysis of solid breast tumour (neoplastic plus stroma) gene expression studies, enabling those genes of specific epithelial origin to be identified in respect to progression, prediction of outcome and metastasis. The expression data obtained from the normal luminal and myoepithelial cells have extended our previous analysis of these normal cell types [11], and provide gene sets that can be used to comprehensively specify the epithelial phenotype expressed in breast tumours, as well as defining new markers of each cell type.

The data presented here report for the first time the application and validation of the MPSS sequencing technology to malignant human breast epithelial cells and their normal counterparts. MPSS expression studies of different human cell lines and normal tissues have already shown that this technology represents the most comprehensive sequencing methodology available at present, in terms of gene coverage and quantitative assessment of gene expression [22,39]. With over 10^6 sequencing reactions per sample [18,19], it is comparable in scope with the now commonly used genome-wide microarray profiling methods, as also used in the present study. Comparative studies of genome wide data sets are entirely dependent on the choice of common denominator for annotation [40]. By using our sequence based mapping, 97% of MPSS tags could be aligned with individual features on genome-wide microarrays, indicating that the vast majority of the expressed sequence tags in the normal and malignant breast epithelium MPSS libraries represent known transcripts, in agreement with the recent data suggesting that MPSS identifies very few truly novel genes [39]. Given the significant methodological differences between microarray and MPSS analysis, the fact that more than 65% of our MPSS differential data set showed concordance with expression profiling obtained by several different microarray platforms, represents a good overlap compared with other examples of sequence versus array data [41]. However, a substantial number of differentially expressed genes (4,149) measured on at least two microarray platforms were not identified as such by MPSS, and a significant number of MPSS differential transcripts (2,440) were not confirmed on any array (Figure 1), implying a relatively high false positive and false negative rate of the MPSS methodology. This probably reflects the known limitations of the MPSS technology [39], particularly with regards to transcripts that were not detected (zero counts) in one sample, as well as genes lacking appropriate restriction enzyme sites required for this technology. However, individual microarray platforms themselves differ substantially [42] and a multiplatform approach, as used here, clearly defines a robust DTET seen by every technology.

Figure 5



Immunohistochemical analysis of periostin (POSTN), IL8 and cartilage oligomeric matrix protein (COMP). **(a)** POSTN-positive invasive ductal carcinoma (IDC; ×400), in which both epithelial and stromal cells show cytoplasmic expression. **(b)** POSTN-negative IDC in which only the spindle shaped stromal cells are stained (×400). **(c)** IL8 (×100), showing positive staining only in the malignant breast epithelial cells. **(d)** COMP expression in the epithelial and stromal cells of an IDC, showing strong expression in both stromal and epithelial cells (×100).

Another important feature of our DTET is the use of purified epithelial cells, derived by both positive and negative immunomagnetic sorting in which the contamination of malignant samples with stromal cells is reduced to a minimum, and normal luminal and myoepithelial cells are separated from short-term primary cultures. Although the profiling techniques used represent the global transcriptomes of purified normal and neoplastic breast epithelial cells in highly enriched preparations, it is conceivable that even a small contamination of the