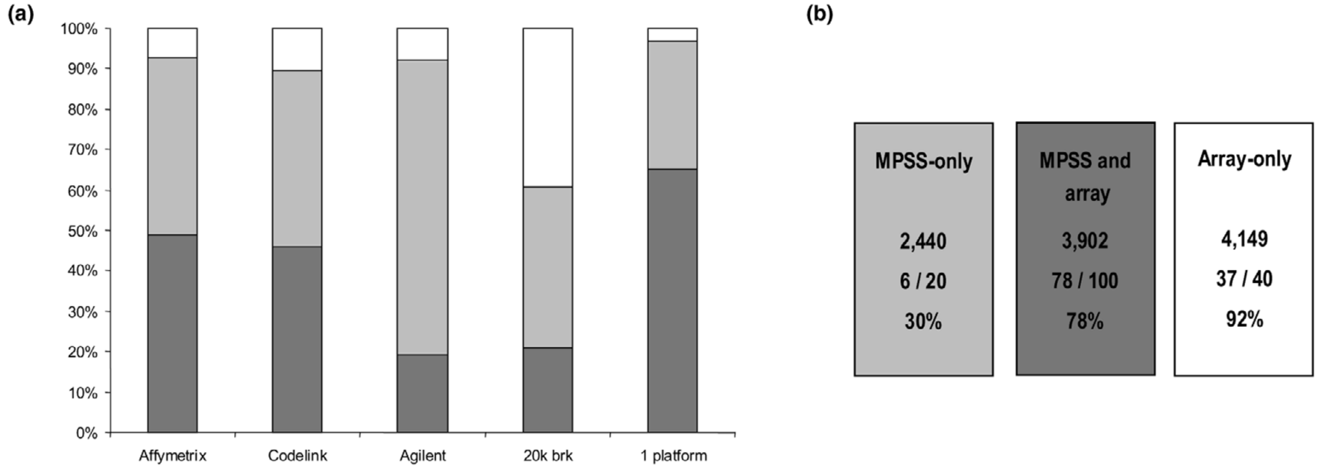
Figure 1



Comparison of massively parallel signature sequencing (MPSS) data with microarray analysis. Differentially expressed gene profiles from MPSS (100%) were overlaid with each microarray platform individually. **(a)** Percentage of coverage (light grey) and concordance in differential expression between MPSS and individual arrays (dark grey) are shown together with the combined coverage and confirmation by at least one array (1 platform). **(b)** Enumeration of the differentially expressed transcripts detected by "MPSS-only", by "MPSS and array", and those transcripts reported as differential by at least two arrays, but not by MPSS ("Array only"). The results obtained by RT-PCR for these subgroups are shown below (see Additional file 6).

Microarray analysis

The same total RNA pools were hybridised onto a 20 k cDNA microarray (20 k brk, constructed at The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK containing 19,391 sequence-validated IMAGE clones), Affymetrix Human Genome U133 Plus 2.0 GeneChip (Affymetrix, Inc., Santa Clara, CA, USA), CodeLink™ Human Whole Genome Bioarray (GE, Healthcare, formerly Amersham Biosciences, Chandler, AZ, USA) and Agilent Whole Human Genome Oligo Microarray 44 k cDNA array (Agilent Technologies, Palo Alto, CA, USA). Three technical replicates of each RNA pool were amplified, labelled and hybridised according to manufacturer's guidelines. Where necessary an RNA pool consisting of breast cancer cell lines was used as a reference sample [11] and dye-swap hybridisations were performed. All primary array data are available through ArrayExpress [20]; they comply with MIAME standards, with the accession number E-TABM-66. Overlay of each microarray platform with MPSS was done by mapping the sequence information of probes and probe sets to the same HTR database as used for MPSS tag mapping (see above). Only those microarray features that were unambiguously mapped to a single HTR cluster were included for further studies. All preprocessing of each microarray platform and further statistical analysis was performed in the R 2.1.1 environment [24] by making extensive usage of the limma package [25] in BioConductor 1.6 [26]. For the Affymetrix platform, probe-level data were normalised and expression data were summarised by the robust multi-array analysis [27]; cyclic lowess normalisation was applied to the CodeLink™ expression data through the *codelink* 0.7.2 package in R 2.3; for the Agilent microarrays, global normalisation with no background correction was applied; and for the 20 k brk microarrays, raw expression data were print-tip normalised and background corrected. Relative measurements for each transcript were given as a $\log_2$ fold ratio, and only genes with a false discovery prediction of $P \leq 0.05$ were regarded as significantly differentially expressed when using Benjamini and Hochberg' s *P* values adjustment [28].

Gene Ontology

Genes were categorised with respect to their biological process, cellular role, molecular function, using Onto-Express (OE) [29,30]. The most significant perturbed biological processes were determined with respect to the number of genes expected for each Gene Ontology (GO) category based on their representation on the Affymetrix U133 Plus 2.0 array. Statistical significance was determined by using OE's hypergeometric probability distribution and Bonferroni correction options, and annotations with $P \leq 0.05$ were accepted as significant. Gene set enrichment analysis (GSEA) comparing luminal and myoepithelial gene signatures was done using described methods [31]. Biological processes were ranked according to their significance of enrichment, and the validation mode measure of significance was used to identify those of greatest enrichment.

Semiquantitative RT-PCR

Total RNA (10 μg) from the normal luminal epithelial and the malignant epithelial RNA pool was used for each 40 μl reverse-transcription reaction, and 10 μl of 1/50 diluted cDNA