Table 1

Numerical analysis of massively parallel signature sequencing

	Malignant breast epithelium	Normal luminal epithelium
Sequence signatures	24,288	28,404
Uniquely mapped signatures	14,245	10,249
Unique HTR clusters	8,421 (3,191)[a]	6,477 (1,297)[a]
Dynamic range	<9,808 tpm	<35,847 tpm
Differentially expressed transcripts	4,311 T > L	2,242 L > T

Sequence signatures represent the total number of sequences obtained by massively parallel signature sequencing (MPSS). Uniquely mapped signatures correspond to the total number of human transcriptome clusters identified and retained in the 'gene-centric' annotation. Unique human transcriptome database (HTR) clusters are transcripts that mapped to a single human cluster and had an abundance of ≥3 transcripts per million (tpm) (approximately one transcript/cell). As described in Materials and methods, statistically significantly ($P \leq 0.05$) differentially expressed transcripts were determined and separated into tumour (T) over normal luminal (L) or vice verse, depending on their fold change. [a]Corresponds to HTR clusters found in only one sample.

medium with 2% fetal calf serum for 4 to 6 h with intermittent shaking. After brief settling, the supernatant was spun down, and the pellet resuspended in L-15 medium and passed through a 100 μm mesh filter to remove residual undisaggregated tumour fragments, plus disaggregated 'normal' organoids and ducts as well as lobules and ducts distended with ductal carcinoma *in situ*, leaving only small clusters and single cells. The latter were then reacted with the mouse monoclonal antibody F19 to fibroblast activation protein bound to sheep anti-mouse coated Dynabeads (Dynal, Paisley, UK) using the manufacturer's protocols. Almost all desmoplastic fibroblasts associated with breast cancers express this antigen strongly. Cells attached to beads were removed with a Dynal MP40 magnet; F19-negative cells were then allowed to sediment under unit gravity for 2 to 3 h (to remove most lymphocytes). The resulting preparation was then screened by phase contrast microscopy to identify those preparations in which there were few if any microvessels (the other main potential stromal contaminant not removed by fibroblast activation protein sorting), or normal tissue elements, such as ducts or acini's. Of the 50 samples, 15 were selected for this study, based on the criteria of ≥80% malignant cell content as determined by phase-contrast examination, ≥80% viability (as determined by trypan blue exclusion) and the integrity of its total RNA. The purity of both normal and malignant epithelial preparations is illustrated in Additional file 1. Informed consent to use this material for scientific research was obtained, and details of the pathology of the individual tumours are given as Additional file 2. RNA was prepared from individual samples by standard Trizol methods and pooled to give a luminal, a myoepithelial and a malignant RNA sample of >1 mg for analysis.

MPSS analysis

MPSS was performed by Lynx Therapeutics, (CA, USA) according to the Megaclone 'signature' protocol [18,19]. Briefly for each library synthesis, after DNase treatment of approximately 300 μg total RNA from normal luminal and malignant breast epithelial pools, cDNA was generated from poly(A)+ RNA, and amplified copies of each cDNA clone were attached to beads. The sequence adjacent to the poly(A) proximal *Dpn*II site was determined by cycles of ligations to fluorescently tagged 'decoding' oligonucleotides and cleavages by restriction enzymes. Each sequence signature comprises the *Dpn*II restriction recognition site (GATC) and 13 contiguous nucleotides. The raw data resulted from four sequencing runs, collected in two reading frames offset by two nucleotides relative to the anchoring restriction enzyme site and generating approximately 2 to 3 × 10^6 sequences. Signatures that were seen in at least two independent runs (reproducible) and were present at a frequency of more than three transcripts per million (tpm) in one sample (significant) were selected for further analysis.

As a basis for the matching of signature sequences to transcripts, we used our own reconstitution of the human transcriptome database (HTR) [21-23] based on a comprehensive set of cDNA to genome alignments that are merged into gene models representing the detailed structure of human transcribed regions. Each HTR contains a cluster of cDNA sequences, similarly to the NCBI/UniGene database. The annotation of the signature was then performed in two steps as described previously [22], using the NCBI35 assembly of the human genome. Firstly, a 'signature-centric' annotation was performed, where sequence signatures were mapped to either one or more transcribed regions of the genome, including repetitive sequences, ribosomal, mitochondrial and non-mapped transcripts. In the second step, only signatures from the 'signature-centric' annotation that matched exactly or had one nucleotide mismatch to known transcribed regions were retained to form the 'gene-centric' version. When different sequence signatures mapped to the same gene, counts were combined. To identify genes with significant differences (P value ≤ 0.05) in representation in the two RNA pools, the absolute difference in abundance between the malignant and the normal epithelial RNA sample was determined and $\log_2$ transformed, resulting in a relative expression measurement.