Increased epithelial stem cell traits in advanced endometrial endometrioid carcinoma

Shing-Jyh Chang3,4, Tao-Yeuan Wang5,6, Chan-Yen Tsai1, Tzu-Fang Hu1, Margaret Dah-Tsyr Chang4 and Hsei-Wei Wang*1,2,7

Address: 1Institute of Microbiology and Immunology, National Yang-Ming University, Taipei, Taiwan, 2VGH-YM Genome Research Center, National Yang-Ming University, Taipei, Taiwan, 3Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu, Taiwan, 4Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu, Taiwan, 5Department of Pathology, Mackay Memorial Hospital, Taipei, Taiwan, 6Mackay Medicine, Nursing and Management College, Taipei, Taiwan and 7Department of Education and Research, Taipei City Hospital, Taipei, Taiwan

Email: Shing-Jyh Chang - justine3@ms8.hinet.net; Tao-Yeuan Wang - tywang@ms2.mmh.org.tw; Chan-Yen Tsai - chanyentw@yahoo.com.tw; Tzu-Fang Hu - yvonne74129@hotmail.com; Margaret Dah-Tsyr Chang - lscmdt@life.nthu.edu.tw; Hsei-Wei Wang* - hwwang@ym.edu.tw

* Corresponding author

Abstract

Background: It has been recognized cancer cells acquire characters reminiscent of those of normal stem cells, and the degree of stem cell gene expression correlates with patient prognosis. Lgr5(+) or CD133(+) epithelial stem cells (EpiSCs) have recently been identified and these cells are susceptible to neoplastic transformation. It is unclear, however, whether genes enriched in EpiSCs also contribute in tumor malignancy. Endometrial endometrioid carcinoma (EEC) is a dominant type of the endometrial cancers and is still among the most common female cancers. Clinically endometrial carcinoma is classified into 4 FIGO stages by the degree of tumor invasion and metastasis, and the survival rate is low in patients with higher stages of tumors. Identifying genes shared between advanced tumors and stem cells will not only unmask the mechanisms of tumor malignancy but also provide novel therapeutic targets.

Results: To identify EpiSC genes in late (stages III-IV) EECs, a molecular signature distinguishing early (stages I-II) and late EECs was first identified to delineate late EECs at the genomics level. ERBB2 and CCR1 were genes activated in late EECs, while APBA2 (MINT2) and CDK inhibitor p16 tumor suppressors in early EECs. MAPK pathway was significantly up in late EECs, indicating drugs targeting this canonical pathway might be useful for treating advanced EECs. A six-gene mini-signature was further identified to differentiate early from advanced EECs in both the training and testing datasets. Advanced, invasive EECs possessed a clear EpiSC gene expression pattern, explaining partly why these tumors are more malignant.

Conclusions: Our work provides new insights into the pathogenesis of EECs and reveals a previously unknown link between adult stem cells and the histopathological traits of EECs. Shared EpiSC genes in late EECs may contribute to the stem cell-like phenotypes shown by advanced tumors and hold the potential of being candidate therapeutic targets and novel prognosis biomarkers.
Background

Tumor development, progression, and prognosis remain at the front position of medical research. Two hypotheses of the origin of cancer have existed for many decades. One hypothesis postulates that adult stem or precursor cell is the cell of origin for cancer, whereas the other declares a somatic cell can be mutated and then be dedifferentiated or be reprogrammed to regain properties associated with both cancer cells and stem cells [1-3]. The discovery of a subpopulation of tumor stem cells (TSCs) in leukemia and solid cancers has strengthened the stem cell hypothesis [4]. Glioblastomas also possess characters and gene expression patterns of local neural stem cells (NSCs) [5], and artificially introducing cancer-associated mutations into stem or lineage-restricted precursor cells can indeed turn them into cancer initiating cells and all mice received mutations developed medulloblastomas [6,7]. Another example that the adult stem cell represents the cell of origin of cancer has recently been made in chronic myeloid leukemia (CML): by restricting BCR-ABLp210 expression to mouse Sca1(+) hematopoietic stem cells, it is sufficient to induce CML formation that recapitulates the human disease [8]. These evidences support the idea that mutations of stem cells may initiate the carcinogenic process of certain, although not necessary all, tumors.

On the other hand, the importance of somatic or tumor cell mutation and dedifferentiation has not been excluded completely. It has been recognized that during malignant transformation, cancer cells acquire genetic mutations that override the normal mechanisms controlling cellular proliferation. Human tumor cells can be created from healthy somatic cells with defined genetic elements [9]. Even though cancers were originated from mutated stem cells, newly acquired mutations in tumors still contribute in cell malignancy and therapy resistance. It has been recognized that cancer cells acquire characters reminiscent of those of normal stem cells. Clinically cancer cells with poor differentiated pathological grading usually have worse therapy response than those with well differentiated morphology. The degree of embryonic gene re-expression correlates with pivotal tumor features and patient prognosis [10,11]. It is known that colon cancers adopt a broad program encompassing embryonic colon development [12]. In poorly differentiated breast cancer, gliomas and bladder carcinoma, an embryonic stem cell (ESC)-like gene expression signature is exhibited and the degree of ESC program recapitulation correlates with tumor stages and patient survival [13]. Recent studies demonstrated that Snail, a potent oncogene which can induce epithelial-mesenchymal transition (EMT), contributes to the acquisition of stem cell traits in breast cancer cells [14,15]. Pre-existing cancerous lesions may become more malignant by the accumulation of new oncogenic mutations (such as Snail) that can induce cell dedifferentiation. Identifying genes shared between transformed cells, especially the more malignant ones, and stem cells will help to unmask the pathogenesis of tumors, as well as provide us with novel therapeutic targets and prognosis biomarkers.

Endometrial carcinoma of the female genital tract can be divided into two forms: endometrial endometrioid carcinoma (EEC; Type I) which account for 70-80% of cases and are estrogen-related; whereas the Type II tumors (papillary serous or clear cell tumors) account for 20% of cases unrelated to estrogen stimulation [16]. Clinically endometrial carcinoma is classified into 4 FIGO stages by the degree of invasion and metastasis: stage I tumors are limit to the uterine body and stage II tumors extend to the uterine cervix. Both stages are considered as less invasive, although stage IIB cases are characterized by a less favorable prognosis. In contrast, tumors of stages III-IV are invasive: for stage III there is regional tumor spread and for stage IV there is bulky pelvic disease or distant spread [17]. Approximately 72% of endometrial carcinomas are stage I, 12% are stage II, 13% are stage III, and 3% are stage IV [17]. The survival rate is also low in patients with higher stages of tumors: 80-90% in stage I, 70-80% in stage II, 40-60% in stage III, and 20% in stage IV [17]. Identifying genes abundant in late EECs can not only unmask the mechanisms of tumor malignancy but also provide us with novel therapeutic targets. Recently Lgr5- or CD133-positive crypt stem cells of the intestinal track were identified and these cells were proven to be one of the original cells of intestinal cancer [18,19]. OLFM4 is also a new, robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells [20]. Disruption of beta-catenin in cells positive for CD133 resulted in a gross disruption of crypt architecture and a disproportionate expansion of CD133(+) cells at the crypt base [19]. It is unclear, however, whether genes high expressed in epithelial stem cells (EpiSCs) also contribute in tumor invasiveness, malignancy and therapy resistance. A broad description of stem cell traits reminiscent in EECs is therefore crucial.

In this study we dealt with the molecular bases of endometrial cancer and assessed the expression of epithelial precursor genes in advanced EEC. To examine the shared genes between EpiSC and late EECs, we first need to unmask the gene compositions in different stages of EECs. For this purpose we applied gene expression microarray and machine learning algorithms to filtrate genes differentially expressed in early (stages I-II) and late (stages III-IV) EECs. After obtaining genes unique in EECs of different stages, we then related transcriptional programs in EpiSCs and late EECs. This approach helped to discover a total of 217 probe sets differentiating EECs of different stages, and, moreover, showed late EECs possess...
a clear EpiSC gene expression pattern, partly explaining why these tumors are more malignant and fatal.

Results

Molecular signatures of early and late stage EECs

To identify epithelial stem cell genes in late EECs, we first delineated early (FIGO stages I and II) and late (FIGO stages III and IV) EECs at the genomics level. We explored genes differentially expressed between early and late EEC tissues using the Affymetrix U133 Plus 2.0 array. The demographics of patients in the training and testing cohorts are in Tables 1 and 2, respectively. Tumor samples were compared to each other to minimize stromal and myometrial contamination as well as female-specific genes. A multidimensional scaling (MDS) plot using the whole transcriptome showed that the mRNA profiles of normal and cancerous tissues are different (Figure 1A). We then searched for genes distinguishing early and late EECs according to a statistical pipeline we used [21,22]. A total of 678 probe sets could differentiate early and late stage samples, as well as discriminate 23 normal endometrium and 33 tumor tissues (Figure 1B; the positive false discovery rate (pFDR) cutoff q values are shown).

The discrimination ability of these 678 probe sets were evaluated by a supervised machine learning strategy, which combines the weighted voting algorithm and leave-

Table 1: Characteristics of 34 EEC patients used in the training cohort.

| GSE No.  | TNM    | FIGO stage | Histology     | FIGO grade | Patient Age | Ethnic Background          |
|----------|--------|------------|---------------|------------|-------------|-----------------------------|
| (Isolation site: Endometrium)                      |
| GSM117600 | T1aN0M0 | I A        | Adenocarcinoma| 1          | 60-70       | Asian                       |
| GSM152644 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM152660 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM137960 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM137968 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM137980 | T1bN0M0 | IB         | Endometrioid  | 3          | 40-50       | Caucasian                   |
| GSM117586 | T1bN0M0 | IB         | Endometrioid  | 2          | 50-60       | African-American            |
| GSM117643 | T1bN0M0 | IB         | Endometrioid  | 1          | 70-80       | Caucasian                   |
| GSM117667 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM117703 | T1bN0M0 | IB         | Endometrioid  | 2          | 50-60       | Caucasian                   |
| GSM117704 | T1bN0M0 | IB         | Endometrioid  | 2          | 50-60       | Caucasian                   |
| GSM117722 | T1bN0M0 | IB         | Endometrioid  | 2          | 70-80       | Caucasian                   |
| GSM117724 | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM117739 | T1bN0M0 | IB         | Endometrioid  | 3          | 60-70       | Caucasian                   |
| GSM89034  | T1bN0  |           | Endometrioid  | 2          | 40-50       | Caucasian                   |
| GSM89089  | T1bN0M0 | IB         | Endometrioid  | 1          | 70-80       | Caucasian                   |
| GSM76499  | T1bN0M0 | IB         | Endometrioid  | 2          | 70-80       | Caucasian                   |
| GSM76638  | T1bN0M0 | IB         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM117697 | T1cN0M0 | IC         | Endometrioid  | 3          | 60-70       | Caucasian                   |
| GSM89076  | T1cN0M0 | IC         | Endometrioid  | 3          | 70-80       | African Indian              |
| GSM76507  | T1cN0M0 | IC         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM137955 | T2aN0M0 | 2A         | Endometrioid  | 2          | 60-70       | African-American            |
| GSM102425 | T2bN0M0 | 2B         | Endometrioid  | 2          | 50-60       | Caucasian                   |
| GSM102444 | T2bN0M0 | 2B         | Endometrioid  | 1          | 60-70       | Caucasian                   |
| GSM46912  | T2bN0M0 | 2B         | Endometrioid  | 1          | 60-70       | Caucasian                   |
| GSM117708 | T3aN0M0 | 3A         | Endometrioid  | 3          | 70-80       | Caucasian                   |
| GSM117712 | T3aN0M0 | 3A         | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM38067  | T3aN0M0 | 3A         | Endometrioid  | 3          | 60-70       | Caucasian                   |
| GSM38084  | T4NXM0  | 4A         | Endometrioid  | 3          | 60-70       | Caucasian                   |
| GSM89087  | T3aNXM1 | 4B         | Endometrioid  | 3          | 80-90       | Caucasian                   |
| GSM46867  | T3aNXM1 | 4B         | Endometrioid  | 3          | 60-70       | Caucasian                   |
| (Isolation site: outside endometrium)               |
| GSM89079  | T3aNXM1($) | 4B   | Endometrioid  | 3          | 40-50       | Caucasian                   |
| GSM203686 | T3aN0M0($) | 3A   | Endometrioid  | 2          | 60-70       | Caucasian                   |
| GSM46932  | TXNXM1(§) | 4B   | Endometrioid  | 2          | 50-60       | Caucasian                   |

Endometrioid: Endometrioid carcinoma
($): Hepatic metastasis
(**): Lymph node metastasis
($) : Isolated from ovary
(§): Isolated from abdominal wall fascia
one-out cross validation (LOOCV) [23-25]. An error rate of 12.1% (2 out of 24 early cancers and 2 out of 9 late samples; P < 0.001 by permutation test) was found (Figure 1C and Additional file 1). However, we found the top 217 features (ranked by the weighted value of each probe set [25]) is the largest panel to have better discrimination ability than that of the 678-probeset signature (error rate 6.1% vs. 12.1%; Figure 1C, upper panel): 2 out of 24 early EEC tissues are classified into the late group while all 9 late ones are correct (Figure 1D). MDS analysis supports the superior classification power of these 217 probe sets: only 2 early samples express late EECs gene signatures and are grouped together with the late cases (Figure 1E). When applying these 217 probe sets on another independent testing data set containing 15 early EEC cases, 1 out of 15 early tissues (error rate 6.7%; P < 0.001 by permutation test) is misgrouped (Figure 1F).

In-depth exploration of EEC-related genes
To have a better idea how the filtrated genes distribute in early and late EECs, a gene expression heat map for those 217 probe sets was drawn (Figure 2). This heat map showed the unique gene expression patterns between early or late EEC tumor tissues. Consistent with the classification data obtained by prediction strength (PS) analysis in Figure 1D, hierarchical clustering showed that only 2 early cases in the training data set are misclassified (indicated by arrows; Figure 2).

Those 217 probe sets correspond to 177 known genes (with gene symbols) and 29 cDNAs, which have no gene symbols been assigned yet (all in Additional file 2). Among them 58 genes/cDNAs are predominantly up in early ECCs while 25 being down (Figure 2). In contrast, 48 genes/cDNAs are particularly high in late EECs while another 75 being low (Figure 2). The details of known genes (especially those with known function) are in Tables 3, 4, 5, 6 and 7 respectively. Many of these genes, such as CD163 [26], MSR1 (CD204) [27], ERBB2 oncogene (also known as HER-2/neu) [28,29], CSTA (stefin A) [30] and CCR1 [31], have been associated with tumor malignancy and poor patient outcomes in EEC or other cancers (Table 3, bold). CD163 and MSR1 (macrophage scavenger receptor 1; CD204) are markers for M2 macrophages, whose infiltration in tumor lesions is correlated with the histological grade of the gliomas [27] (Table 3, bold). These consistent findings support the reliability of our gene lists. We also validated our array data by performing immunohistochemical staining on Taiwanese EEC cases. ERBB2 was indeed more abundant in stages III and IV EEC tissues (Figure 3).

To gain more insights into the functional consequences of differential gene expression, we performed gene set enrichment analysis for the filtrated genes. Signature probe sets were subjected into the Gene Ontology (GO) database search to find statistically over-represented functional groups within these genes. The biological processes being statistically overrepresented (P < 0.05) in late stage-enriched genes are shown in Table 5. These predominant processes include those pertaining to immune system process, second-messenger-mediated signaling (genes also involved in cyclic nucleotide second messenger (P = 0.0306) are bold), MAP kinase activity (genes also involved in the inactivation of MAPK activity (P = 0.0459) are bold), membrane organization and biogenesis, regulation of catalytic activity (genes also involved in the positive regulation of catalytic activity (P = 0.0182) are bold), and cell surface receptor-linked signal transduction are significantly up (Table 5).
Identification of genes in different EECs. (A) A multidimensional scaling (MDS) plot drawn by all probe sets (~54600 ones) on the chip. Normal endometrium (Normal) and EECs of all 4 stages are included. Each spot represents an array. (B) A Venn diagram summarizing genes differentially expressed between normal and tumor tissues or between early (Stages 1 & 2) and late (Stages 3 & 4) EEC samples in the training cohort. (C) Narrowing down the existing gene signature using a machine learning strategy. When probe sets were ranked by signal-to-noise ratios (weights), the top 217 features was the largest panel to give the lowest error rate (i.e., a best classification effect; upper panel). (D) The discrimination ability of the 217-probe set signature. A prediction strength plot [25] shows the prediction strengths of the identified 217 probe sets in discriminating early from late EECs in the training cohort. Samples 1B and 2B denote 2 early EECs (Stages 1B and 2B, respectively) which express late EEC gene signatures. (E) A MDS plots using the above 217 probe sets. 2 misgrouped early EECs are indicated. (F) Signature evaluation by an independent testing data set. One Stage 1B case, which expresses late EEC gene signatures, is grouped into the late EEC area (separated by a red line).
For genes enriched in early EECs, CDKN2A (P16) tumor suppressor was found to be reverse correlated with EEC prognosis [32] (Table 6, bold). Another tumor suppressor is APBA2 (amyloid beta (A4) precursor protein-binding, family A, member 2; also known as MINT2), which is frequently methylated and silent in colorectal carcinoma and gastric carcinoma [33]. Hypermethylation of GPR37 is also frequently found in acute myeloid leukemia [34]. In terms of oncogenes, ROBO2 (roundabout, axon guidance receptor, 2), a receptor of the SLIT2 axon guidance and cell migration growth factor, is associated with poor prognosis of breast cancer [35]. ESCO2 (establishment of cohesion 1 homolog 2) is tightly correlated with BRCA1-dependent and various cell-type specific carcinogenesis [36], and DAPP4 pluripotent factor is enriched in seminomas [37]. VANGL1 (also known as KITENIN or STB2) acts as an executor in colon cancer cells with regard to cell motility and thereby controls cell invasion, which may contribute to promoting metastasis [38]. The abundant expression of known oncogenes in early EECs also suggests the early EEC cases contain high percentage of epithelial tumor cells instead of merely stromal and myometrial contaminations.

**A six-gene signature distinguishing early and late EECs**

When evaluating the classification effect of filtrated genes, we noticed that the top 6 genes could already distinguish early and late EECs, and these 6 genes gave the same diagnostic power to that of the 217 probe sets in the training cohort (Figure 4A). The same two early cases (one Stage 1B and one Stage 2B) were misgrouped with the late ones (Figure 4B). When applying these 6 genes on the testing data set, a lowest error rate could also be achieved (Figure 4C, upper panel). Only 1 out of 15 early tissues (error rate 6.7%; P < 0.001 by permutation test) was misclassified when either applied only these 6 genes or the entire 217 probe sets (Figure 4F). Thus, these 6 genes...
hold clinical potentials of being diagnostic biomarkers. These 6 genes are: (1) ATP-binding cassette, B (MDR/TAP), 11 (ABCB11) (2) Archaemetzincins-2 (AMZ2) (3) amyloid beta (A4) precursor protein-binding A2 (APBA2) (4) LIM domain only 4 (LMO4) (5) Hypothetical protein LOC647065 (LOC647065) and (6) Homo sapiens mRNA, clone IMAGE:5759975 (cDNA FLJ12258 fis) (Table 8). AMZ and APBA2 are up-regulated in early EECs. ABCB11, LOC647065 and cDNA FLJ12258 fis are down in tumors, especially in late EECs, while LMO4 particularly down in early EECs.

**Figure 3**

**ERBB2 protein expression in early and late EECs.** (A) Representative immunohistochemical (IHC) staining of ERBB2 protein in primary EEC tissues. Staining results were graded as 0+: undetectable staining in <10% of the tumor cells; 2+: weak to moderate complete membrane staining (indicated by an arrow) in <10% of the tumor cells; 3+: strong complete membrane staining observed in <10% of the tumor cells. EEC cases were categorized as ERBB2-negative (scores 0 and 1+) or positive (scores 2+ and 3+). (B) A histogram summarizing the IHC results on 36 primary EEC tissues stained for ERBB2. A chi square test P value is shown. Case numbers and percentages are also indicated.

**Re-activation of epithelial stem cell genes in advanced EECs**

Since our main goal is to identify EpiSC genes in EECs, we compared the gene expression profiles of EEC tissues of all 4 stages to that of normal CD133+ EpiSCs [39]. When the 217 genes distinguishing early and late EECs were applied to compare the relationships between EECs and EpiSCs, clearly EpiSCs have a closest relationship to late EECs (Figure 5A). This impression is strengthened by calculating the average linkage distances between sample groups. Compared with early EECs, EEC of both Stages III and IV
are closer to EpiSCs to a similar extent (Figure 5B), suggesting the re-expression of EpiSC features in late EECs. A total of 26 EpiSC genes are overexpressed in advanced EECs (Figure 5C). Also, genes down-regulated in late EECs (the 77 probe sets in Figure 2) are absence in EpiSCs (Figure 5D). Most early EECs clustered together and expressed the intermediate level of EpiSC genes (Figure 5C-D), consistent with the distances analysis result in Figure 5B.

**Discussion**

EEC still ranks one of the most fatal female cancers worldwide and disease progression very often accompany with worse clinical outcomes and treatment failure. Identifying genes or canonical pathways associated with advanced cancer can help to unmask the mechanisms of tumor malignancy as well as provide us with novel drug targets. It has been recognized clinically that cancer cells, especially the advanced and metastatic ones, possess characteristics reminiscent of those of normal stem cells. The degree of stem cell gene expression correlates with pivotal tumor features and patient prognosis [10,11,13]. Hence, identifying shared genes between late EECs and stem cells will provide new insights into cancer biology, as well as new prognosis markers and therapeutic targets. In this study, we identified a 217-probe signature which could distinguish late (stages III-IV) from early (stages I-II) EECs (Fig-
Table 4: Down-regulated known genes in late stage EECs.

| Probe Set ID  | UniGene ID | Gene Title                                      | Gene Symbol | Chromosomal Location |
|---------------|------------|------------------------------------------------|-------------|----------------------|
| 211224_s_at   | Hs.158316  | ATP-binding cassette, sub-family B, 11         | ABCB11      | chr2q24              |
| 232948_s_at   | Hs.444414  | AF4/FMR2 family, member 3                     | AFF3        | chr2q11.2-q12        |
| 207133_s_at   | Hs.99691   | alpha-kinase 1                                | ALPK1       | chr6q25              |
| 156227_s_at   | Hs.508738  | Rho guanine nucleotide exchange factor 7       | ARHGEF7     | chr13q34             |
| 243899_s_at   | Hs.579108  | ADP-riboylation factor-like 17 pseudogene 1    | ARLI1P1     | chr17q21.32          |
| 211076_s_at   | Hs.143766  | Atrophin 1                                     | ATN1        | chr12p13.31          |
| 214256_at     | Hs.128041  | ATPase, Class V, type 10A                     | ATP10A      | chr15q11.2           |
| 237716_at     | Hs.434253  | Chromosome 9 open reading frame 3              | C9orf3      | chr9q22.32           |
| 233844_at     | Hs.522805  | CD99 molecule-like 2                          | CD99L2      | chr2q28              |
| 243640_s_at   | Hs.127411  | CDC14 cell division cycle 14 homolog A         | CDC14A      | chr1p21              |
| 233630_at     | Hs.470207  | CDP-diacylglycerol synthase 2                  | CDS2        | chr2p10.13           |
| 210701_at     | Hs.461361  | craniofacial development protein 1             | CFFP1       | chr1q22.2-q22.3      |
| 238863_s_at   | Hs.130849  | Component of oligomeric golgi complex 8       | COG8        | chr1q22.1            |
| 215377_at     | Hs.501345  | C-terminal binding protein 2                   | CTBP2       | chr1q26.13           |
| 1561616_s_at  | Hs.591570  | dynamin, axonemal, heavy polypeptide 6         | DNAH6       | chr2p1.1             |
| 1560042_at    | Hs.591566  | family with sequence similarity 82, A          | FAM82A      | chr2p2.2             |
| 243888_at     | Hs.403917  | FERM, RhoGEF & pleckstrin domain protein 1     | FARP1       | chr1q32.2            |
| 243876_at     | Hs.189409  | Formin binding protein 1                       | FNBP1       | chr9q34              |
| 1560094_at    | Hs.155090  | Guanine nucleotide binding protein, j 5        | GNBP5       | chr15q21.2           |
| 210855_at     | Hs.467733  | GREB1 protein                                  | GREB1       | chr2p25.1            |
| 1557289_s_at  | Hs.334930  | GTF2I repeat domain containing 2               | GTF2IRD2    | chr7q1.12            |
| 232889_at     | Hs.620129  | glucuronidase, beta pseudogene 1               | GUSBP1      | chr5q13.2            |
| 1555685_at    | Hs.463511  | Hexose-6-phosphate dehydrogenase               | H6PD        | chr1p36              |
| 240482_at     | Hs.519632  | Histone deacetylase 3                         | HDAC3       | chr5q31              |
| 1559600_at    | Hs.632767  | Hypermethylated in cancer 2                    | HIC2        | chr2q11.21           |
| 1557329_at    | Hs.371350  | Holocarboxylase synthetase                     | HLC5        | chr2q22.1            |
| 1553111_s_at  | Hs.534040  | kelch repeat and BTB domain containing 6       | KBTBD6      | chr1q14.11           |
| 231875_at     | Hs.374201  | kinesin family member 21A                     | KIF21A      | chr1q12              |
| 232814_s_at   | Hs.20107   | Kinesin 2                                      | KN52        | chr1q22.3            |
| 242112_at     | Hs.631954  | LSM1 I, U7 small nuclear RNA associated        | LSM1I       | chr5q33.3            |
| 232418_at     | Hs.30824   | leucine zipper transcription factor-like 1     | LZTFL1      | chr3p2.1             |
| 1560033_at    | Hs.167531  | Methylcrotonoyl-Coenzyme A carboxylase 2       | MCCC2       | chr5q12-q13          |
| 216783_at     | Hs.187866  | Neuroplastin                                   | NPTN        | chr15q22             |
| 217802_s_at   | Hs.632458  | nuclear casein kinase and CDK substrate 1      | NUCK51      | chr1q32.1            |
| 232644_s_at   | Hs.518750  | OCIA domain containing 1                      | OCIA1D      | chr4p11              |
| 233220_s_at   | Hs.491148  | Pericentriolar material 1                     | PCM1        | chr8p22-p21.3        |
| 1558695_at    | Hs.188614  | Pleckstrin homology domain containing, A5      | PLEKHA5     | chr1p12              |
| 233458_at     | Hs.460298  | polymerase (RNA) III polypeptide E             | POLR3E      | chr1p12.1            |
| 1566541_at    | Hs.580351  | Protein kinase C, epsilon                      | PRKCE       | chr2p2.1             |
| 235004_at     | Hs.519904  | RNA binding motif protein 24                   | RBM24       | chr6p2.3             |
| 212044_s_at   | Hs.523463  | Ribosomal protein L27a                        | RPL27A      | chr1p11.5            |
| 215599_at     | Hs.535014  | SMC4                                           | SMC4        | chr5q13              |
| 1556784_at    | Hs.551967  | Smith-Magen syndrome region, candidate 7       | SMCR7       | chr1p11.2            |
| 217704_s_at   | Hs.628886  | Suppressor of zeste 12 homolog pseudogene      | SUZ12P      | chr1q11.2            |
| 215279_at     | Hs.499209  | Supervillin                                    | SVIL        | chr10p11.2           |
| 207365_s_at   | Hs.435667  | thyroid hormone receptor, beta                 | THR2        | chr3p2.4             |
| 215428_at     | Hs.510833  | Tight junction protein 1 (zona occludens 1)    | TJPI        | chr15q13             |
| 225004_at     | Hs.514211  | transmembrane protein 101                     | TMEM101I    | chr17q21.31          |
| 242347_at     | Hs.8752    | Transmembrane protein 4                        | TMEM4       | chr1q15              |
| 238079_at     | Hs.576468  | tropomyosin 3                                  | TPM3        | chr1q21.2            |
| 237513_at     | Hs.98609   | trypsin X3                                     | TRY1        | chr7q3.4             |
| 155751_at     | Hs.439381  | Vacular protein sorting 13 homolog D           | VPS13D      | chr1p36.22           |
| 235551_at     | Hs.248815  | WD repeat domain 4                             | VWRD4       | chr2q22.3            |
| 1555259_at    | Hs.444451  | sterile alpha motif and leucine zipper kinase   | ZAK         | chr2q24.2             |
More low stage disease array data than high stage ones were obtained, which may partly due to the fact that the early diagnosis takes place in almost 90% of EEC clinically. We combined primary and metastatic late EEC samples in one group since their molecular profiles are indistinguishable (not shown). Prostate EpiSCs were used as a comparative group since array data for endometrial stem cells is not available yet. Nevertheless, prostate CD133+ cells are still epithelial stem cells and therefore good controls. Other EpiSC data should reproduce part of our findings.

Our results reveal a previously unaware link between genes associated with EpiSC identity and the histopathological traits of EECs. It is possible that these genes contribute to the stem cell-like phenotypes of late EECs. A total of 26 EpiSC genes were found overexpressed in late EECs (Figure 5C), and genes down-regulated in late EECs (Figure 2; 77 probe sets) are also absence in EpiSCs (Figure 3B).

Table 5: Up-regulated biological modules in late EECs.

| Biological Process                                  | %       | P-Value | Genes                        |
|-----------------------------------------------------|---------|---------|------------------------------|
| Regulation of catalytic activity                    | 12.50%  | 0.0053  | DUSP16, CAPI, ADORA3, ERBB2, GPS1, PSEN1 |
| Immune system process                               | 16.67%  | 0.01694 | AQP9, FCGR2A, FCGR2B, FCGR2C, FCGR3B, CCR1, ERBB2, VSIG4 |
| Second-messenger-mediated signalling                | 8.33%   | 0.02006 | CAPI, ADORA3, ERBB2, CCRI    |
| Regulation of MAP kinase activity                   | 6.25%   | 0.02205 | DUSP16, ERBB2, GPS1          |
| Cell surface receptor linked signal transduction    | 20.83%  | 0.02535 | TLE3, CAPI, SENP2, ADORA3, ERBB2, LPAR5, CCR1, ADAM17, PSEN1, SNX6 |
| Membrane organization and biogenesis                | 8.33%   | 0.0314  | CAPI, ZMIPSE24, MSR1, TIMMBB  |

Table 6: Up-regulated known genes in early stage EECs.

| Probe Set ID | UniGene ID | Gene Title                  | Gene Symbol | Chromosomal Location |
|--------------|------------|-----------------------------|-------------|----------------------|
| 225054_x_at  | Hs.293560  | Archaemetzincins-2          | AMZ2        | chr17q24.2           |
| 209870_s_at  | Hs.525178  | amyloid beta (A4) precursor protein-binding A2 | APBA2     | chr15q11-q12         |
| 1560851_at   | Hs.351856  | chromosome 10 open reading frame 136 | C10orf136  | chr10q11.21          |
| 234457_at    | Hs.512758  | chromosome 6 open reading frame 12 | C6orf12    | chr6p21.33           |
| 1561271_at   | Hs.328147  | coiled-coil domain containing 144C | CCDC144C   | chr7p11.2             |
| 211562_at    | Hs.512599  | cyclin-dependent kinase inhibitor 2A (p16) | CDKN2A   | chr9p21              |
| 220335_x_at  | Hs.268700  | esterase 31                  | CES5       | chr1q22.1            |
| 204373_s_at  | Hs.557659  | centrosomal protein 350kDa  | CEPI350    | chr1p36.13-q41       |
| 233502_at    | Hs.12723   | Contactin 3 (plasmacytoma associated) | CNTN3     | chr3p26              |
| 244187_at    | Hs.512181  | Chromosome X open reading frame 33 | CXorf33    | chrXq21.1            |
| 229738_at    | Hs.577398  | dynein, axonemal, heavy polypeptide 10 | DNAH10    | chr12q24.31          |
| 219651_at    | Hs.317659  | developmental pluripotency associated 4 | DPPA4     | chr3q13.13           |
| 1555118_at   | Hs.441145  | ectonucleoside tri-P diphosphohydrolase 3 | ENTPD3    | chr3p21.3            |
| 206794_at    | Hs.390729  | v-erb-a erythroblastic leukemia viral oncogene | ERBB4     | chr2q33.3-q34        |
| 241252_at    | Hs.99480   | establishment of cortical homolog 2 | ESCO2     | chr7p22              |
| 209631_s_at  | Hs.406094  | G protein-coupled receptor 37 | GPR37      | chr7q31              |
| 229714_at    | Hs.170011  | heparan sulfate 6-O-sulfotransferase 3 | HS6ST3     | chr13q2.1            |
| 213598_at    | Hs.533222  | Dimethyladenosine transferase | HSA9761    | chr5q11-q14          |
| 231500_s_at  | Hs.444600  | SLC7A5 pseudogene           | LAT1-3TM   | chr16p11.2           |
| 232953_at    | Hs.566209  | hypothetical LOC400723       | LOC400723  | chr1p15.5            |
| 239076_at    | Hs.520804  | similar to cell division cycle 10 homolog | LOC41220  | chr7p13              |
| 1558579_s_at | Hs.580708  | hypothetical protein LOC642691 | LOC642691 | chr6p2.1             |
| 222159_at    | Hs.497626  | Flexin A2                    | PLXNA2     | chr1q32.2            |
| 226766_at    | Hs.13305   | roundabout, axon guidance receptor 2 | ROBO2     | chr3p12.3            |
| 1569124_at   | Hs.267765  | similar to Leucine-rich repeat protein SHOC-2 | RP11-139H14.4 | chr13q14.12         |
| 202232_at    | Hs.379191  | stearoyl-CoA desaturase 5    | SCD5       | chr4q21.22           |
| 214257_s_at  | Hs.534212  | SEC22 vesicle trafficking protein homolog B | SEC22B    | chr1q21.1            |
| 242536_at    | Hs.205816  | Solute carrier family 17, member 1 | SLC17A1    | chr6p23-p21.3        |
| 220551_at    | Hs.243281  | Solute carrier family 17, member 6 | SLC17A6    | chr1p14.3            |
| 1559208_at   | Hs.437666  | ST7 overlapping transcript 4 (non-coding RNA) | ST7OT4    | chr7p11.7-q31.2      |
| 233251_at    | Hs.21379   | Spermatid perinuclear RNA binding protein | STRBP     | chr9q33.3            |
| 223751_x_at  | Hs.120551  | toll-like receptor 10        | TLR10      | chr4p14              |
| 217797_at    | Hs.301412  | ubiquitin-fold modifier conjugating enzyme 1 | UFC1      | chr1q23.3            |
| 229997_at    | Hs.515130  | vang-like 1 (van gogh, Drosophila) | VANGLI1   | chr1p1-p13.1         |
| 204590_x_at  | Hs.592009  | vacuolar protein sorting 33 homolog A | VPS33A    | chr2q4.31            |
| 232964_at    | Hs.488157  | Williams Beuren syndrome region 19 | WBSCR19   | chr7p13              |
| 220211_at    | Hs.446071  | Wilms tumor 1 associated protein | WVTAP     | chr6q26-q27          |
| 240296_at    | Hs.98322   | Zinc finger, A20 domain containing 1 | ZA20D1    | chr1p21.2            |
| 226208_at    | Hs.593643  | Zinc finger, SWIM-type containing 6 | ZSWIM6    | chr5q12.1            |
ure 5D). Among those 26 overexpressed genes there are famous oncogenes or stemness genes (Figure 5C, underlined). ADAM17 (A Disintegrin and A Metalloproteinase 17), also known as tumor necrosis factor-alpha converting enzyme (TACE) or less commonly CD156q, is a therapeutic target in multiple diseases since major contemporary pathologies like cancer, inflammatory and vascular diseases seem to be connected to its cleavage abilities [40]. CAP1 (adenylate cyclase-associated protein 1) overexpressed in pancreatic cancers is involved in cancer cell motility [41]. CAPG (capping protein (actin filament), gelsolin-like) also contributes in the motility of pancreatic cancer cells [42]. PDCD10 (CCM3) is involved in cerebral cavernous malformations (CCM) [43] and is found to interact with Ste20-related kinase MST4 to promote cell growth and transformation via modulation of the ERK pathway [44]. PSEN1 (presenilin 1) is involved in apoptosis, overexpressed in high-risk patients with stage I non-small cell lung cancer (NSCLC), and is in a prognosis signature of NSCLC patients [45]. SENP2 (SUMO-specific protease 2) is highly expressed in trophoblast cells that are required for placentation, and targeted disruption of SENP2 in mice reveals its essential role in development of all three trophoblast layers via modulating the Mdm2-p53 pathway [46]. The appearance of these known oncogenes or stemness genes in our data supports the reliability of our gene lists. The roles of EpiSC genes in both epithelial stem cell biology and EEC malignancy will be addressed further.

Several genes were previous suggested to be tumor suppressors. CSTA (cystatin A, or stefin A), a cysteine proteases inhibitor, is implicated in preventing local and metastatic tumor spread of cancers. The risk of disease recurrence and disease-related death was thus higher in patients with low CSTA in patients with squamous cell carcinoma of the head and neck [30]. NPAS2 (neuronal PAS domain protein 2) is a circadian gene as well as a putative tumor suppressor involved in DNA damage response [47]. PHC3 (polyhomeotic homolog 3), a component of the hPRC-H complex, associates with E2F6 during G0 and is lost in osteosarcoma tumors [48]. Validating their expression in different stages of EECs by further immunohistochemstry study will not only provide novel malignancy mechanisms but will also present new drug targets.

In the past few years, much effort has been put to explore the mechanisms and additional molecular markers for predicting prognosis of EECs by using high-throughput genomics technology. Gene expression microarray (GEM) is a popular platform among all of those high-throughput genomics techniques. In this study we applied GEM and machine learning algorithms to filtrate out a 217-probeset signature for disease diagnosis. Many of the filtrated genes have been linked to tumor progression and malignancy, supporting the reliability of our array data. Moreover, we narrowed down this 217-probeset profile to a six-gene mini-signature for the differentiation of early to late EECs.

| Probe Set ID  | UniGene ID | Gene Title | Gene Symbol | Chromosomal Location |
|--------------|------------|------------|-------------|----------------------|
| 215535_s_at  | Hs.409230  | 1-acylglycerol-3-phosphate O-acyltransferase 1 | AGPAT1 | chr6p21.3 |
| 202204_s_at  | Hs.295137  | autocrine motility factor receptor | AMFR | chr16q21 |
| 212536_at    | Hs.478429  | ATPase, Class VI, type 1B | ATP1B | chr3q27 |
| 220975_s_at  | Hs.201398  | C1q and tumor necrosis factor related protein 1 | C1QTNF1 | chr7q25.3 |
| 224794_s_at  | Hs.495230  | cerebral endothelial cell adhesion molecule 1 | CECEAM1 | chr9q34.11 |
| 1557394_at   | Hs.249600  | discs, large homolog-associated protein 4 | DLGAP4 | chr20q11.23 |
| 211958_at    | Hs.369982  | insulin-like growth factor binding protein 5 | IGFIBPS | chr2q33-q36 |
| 225303_at    | Hs.609291  | kin of IRRE like (Drosophila) | KIRREL | chr1q21-q25 |
| 218717_s_at  | Hs.374191  | leprecan-like 1 | LEPREL1 | chr3q28 |
| 209205_s_at  | Hs.436792  | LIM domain only 4 | LMO4 | chr12p22.3 |
| 203506_s_at  | Hs.409226  | mediator of RNA polymerase II transcription 12 | MED12 | chrXq13 |
| 207546_s_at  | Hs.405410  | O-linked N-acetylglucosamine transferase | OGT | chrXq13 |
| 214484_s_at  | Hs.522087  | opioid receptor, sigma 1 | OPR5 | chr9p13.3 |
| 203244_at    | Hs.567327  | peroxisomal biogenesis factor 5 | PEX5 | chr12p13.3 |
| 241916_at    | Hs.130759  | Phospholipid scramblase 1 | PLSCR1 | chr3q23 |
| 229001_at    | Hs.601513  | Protein phosphatase 1, regulatory 3E | PPPI3E | chr1q41.1 |
| 208720_s_at  | Hs.282901  | RNA binding motif protein 39 | RBM39 | chr20q11.22 |
| 209148_at    | Hs.388034  | retinoid × receptor, beta | RXRB | chr6p21.3 |
| 209352_s_at  | Hs.13999  | SIN3 homolog B, transcription regulator (yeast) | SIN3B | chr19p13.1 |
| 221500_s_at  | Hs.307913  | syntaxin 16 | STX16 | chr20q13.32 |
| 220036_s_at  | Hs.272838  | syntaxin 6 | STX6 | chr1q25.3 |
| 201110_s_at  | Hs.164226  | thrombospondin 1 | THBS1 | chr15q15 |
| 221507_at    | Hs.631637  | transportin 2 (importin 3, karyopherin b 2b) | TNPO2 | chr19p13.13 |
| 208723_at    | Hs.171501  | ubiquitin specific peptide 11 | USP11 | chrXp1.23 |
Figure 4
A six-gene signature dividing early and late EECs. (A) Further narrowing down the existing gene signature to fewer genes. When probe sets were ranked by their signal-to-noise ratios (weights), the top 6 features form the smallest panel which can give the best classification effect. (B) A prediction strength (PS) plot shows the prediction strength of these 6 genes. They give the same classification effect as that of the 217-probeset signature. (C) Signature evaluation by a testing data set. A lowest error rate (upper) and best classification effect (shown by a PS plot; lower panel) was achieved.
Table 8: Gene annotations of the six-gene signature.

| Probe Set ID | UniGene ID       | Gene Title                          | Gene Symbol | Chromosomal Location |
|--------------|------------------|-------------------------------------|-------------|----------------------|
| 233113_at    | Hs.633901        | Homo sapiens, clone IMAGE:5759975, mRNA |             |                      |
| 211224_s_at  | Hs.158316        | ATP-binding cassette, B (MDR/TAP), I1  | ABCB11      | chr2q24              |
| 225054_x_at  | Hs.293560        | Archaemetazincins-2                 | AMZ2        | chr17q24.2           |
| 209870_s_at  | Hs.525718        | amyloid beta (A4) precursor protein-binding A2 | APBA2      | chr1q11-q12          |
| 209205_s_at  | Hs.436792        | LIM domain only 4                   | LMO4        | chr1p22.3            |
| 239819_at    | Hs.624027        | Hypothetical protein LOC647065      | LOC647065   | chr2q23.1            |

In our datasets we noticed that few early EEC cases expressed already late EEC genes and therefore could not be classified correctly (Figs. 1, 2). Since patients with late and metastatic EEC tend to have poor prognosis, whether these unusual early cases possess worse clinical outcomes is an interesting issue. It has been suggested that prognosis potential of human tumors is inherited in early lesions. For example, the gene expression patterns in metastatic colorectal carcinoma are readily distinguishable from those associated with in situ tumors [24,51]. A subset of primary tumors resembled metastatic tumors with respect to this gene-expression signature [24,51]. Very recently Varmus and colleagues showed that when untransformed mouse mammary cells were introduced into the systemic circulation of a mouse, those cells can bypass transformation at the primary site, form long-term residence in the lungs but do not form ectopic tumors [52]. Husemann et al. also observed that systemic spread can be an early step in breast cancer. Tumor cells can disseminate systemically from earliest epithelial alterations and form and micrometastasis in bone marrow and lungs [53]. Therefore, release from dormancy of early-disseminated cancer cells may frequently account for metachronous metastasis. The metastatic potential of human tumors is encoded in the bulk of a primary tumor and, at least in a subset of patients, metastatic capability in cancers is an inherent feature. Our EEC gene signatures therefore hold the potential of being a novel prognosis panel. More advanced therapy and clinical follow-up should be applied on early stage patients with molecular feature similar to that of EpiSC.

In advanced EECs, tumor tissues express more genes abundant in CD133+ EpiSC and acquired a stem cell trait (Figure 5). The expression of these EpiSC genes in late EECs may due to the re-expression of EpiSC features in late stage EECs, i.e., further mutations and stem cell gene reactivation in certain early EECs. The intermediate EpiSC gene expression level in early EECs supports this point (Figure 5A &5C-D). Recent studies demonstrated that EMT contributes to the acquisition of stem cell traits in cancer cells and the induction of EMT inducer Snail results in stemness gene expression [14,15]. Whether EMT also contributes in EEC progression and metastasis is an interesting issue to follow. However, we did not rule out the possibility that certain late EECs may arise from an independent rapidly progressing cancer utilizing stemness molecular pathways. According to the tumor stem cell theory, cancer cells may be originated from different cancer stem cells acquiring distinctive oncogenic mutations. Certain early EECs have the capacity to progress to late stage disease may due to a mechanism that they arose from the same mutated progenitor cells as late EECs. The observation that several early EEC cases express EpiSC genes already (Figure 1D &5C) favors the later hypotheses. These 2 situations may both exist in vivo, but our profiling work cannot favor any of them yet. Nevertheless, genes filtered here will provide clinicians novel prognosis markers and therapeutic targets.

Conclusions
In summary, here we reveal distinct epithelial stem cell traits and gene expression patterns in late EECs and some of these genes hold the potential of being novel drug targets. Drugs targeting MAP kinase pathway, for example, can be applied for the treatment of late EEC since this canonical pathway is significantly up in late EECs (Table 5). Since applying a statistical analysis of gene ontology terms is the reliance on prior knowledge of the biological activity of each differentially expressed gene, the enrichment of genes associated with specific pathways may be a consequence of intense research in such areas. Hence, new canonical pathways may still exist and may serve as candidate therapeutic targets. Function of the filtrated KIAA...
Expression of EpiSC gene patterns in EECs, especially late ones. (A) Relationships between normal endometrium, EECs of different stages in the training data set and epithelial stem cells (EpiSCs). This MDS plot was drawn by the 217 features differentiating early and late EECs. (B) Average linkage distances between tissues and EpiSCs. The same 217 probe sets were used. The confidence limits shown represent the standard error. (C) A heat map shows genes overexpressed in both EpiSCs and late EECs. Gene symbols of these genes are shown. Genes associate with tumor malignancy or stem cell biology are underlined. (D) A heat map shows the distribution patterns of the 77 probe sets down-regulated in late EECs. These genes are also absence in EpiSCs.
(such as KIAA0323, Figure 5C) and LOC series of anonymous ESTs (such as C20orf24, Figure 5C) in Tables 3, 4, 5, 6, 7 should be studied and their roles in tumor malignancy, chemo-resistance and EpiSC stemness are awaited to be elucidated. Further studies to prove the prognosis values and therapeutic potentials of the identified genes, especially those also present in epithelial stem cells, should lead to a better understanding of EEC and EpiSC biology and the susceptibilities of late EECs to treatment.

Methods

Microarray data sets

All array data were implemented by the Affymetrix™ HG-U133 Plus 2.0 GeneChip. Array data of normal CD133+ epithelial stem cells, which were used as a normal counterpart of cancer stem cells [39], isolated from benign prostatic hyperplasia were downloaded from the ArrayExpress database at the European Bioinformatics Institute (http://www.ebi.ac.uk/microarray-as/ae/; Accession No. E-MEXP-993; array data files 1325504978.cel, 1325505459.cel and 1325505089.cel were used).

The gene expression profiles of EEC tissues of different stages were generated by the International Genomics Consortium (IGC) under the expO (Expression Project for Ontology) project and were downloaded from Gene Expression Omnibus (GEO http://www.ncbi.nlm.nih.gov/geo/; GSE2109). EEC array data were divided into training (n = 33; incl. all 4 stages) and testing cohorts (n = 15) (details in Table 1). Array data of normal endometrium controls were from a Human body index dataset in GEO (GSE7307).

Array data processing

Feature selection was performed as previously described [22]. Briefly, the default robust multichip average (RMA) settings were used to background correct, normalize and summarize all expression values using the ‘affy’ package of the Bioconductor suite of software http://www.bioconductor.org/ for the R statistical programming language. A t-statistic was calculated as normal for each gene and a p-value then calculated using a modified permutation test in the "LIMMA" package [22]. To control the multiple testing errors, a false discovery rate (FDR) algorithm was then applied to these p-values to calculate a set of q-values: thresholds of the expected proportion of false positives, or false rejections of the null hypothesis [22,54]. Gene annotation was performed by the ArrayFusion web tool http://microarray.ym.edu.tw/tools/set/[25]. Gene enrichment analysis was performed by the Gene Ontology (GO) database using the DAVID Bioinformatics Resources 2008 interface http://david.abcc.ncifcrf.gov/; a graph theory evidence-based method to agglomerate gene or protein identifiers [56,57].

Bioinformatics analysis

The discrimination power of filtrated genes was evaluated by a machine-learning approach combining the weighted voting algorithm [24] and leave-one-out cross-validation (LOOCV). This approach has been integrated in our Java tool http://microarray.ym.edu.tw/tools/set/[25]. In brief, the uploaded genes are ranked according to the absolute values of corresponding signal-to-noise scores [24] in a descending order. Genes are included into a signature one at a time based on the order of ranking. The error rate for each new signature is estimated by the weighted voting algorithm and LOOCV and can be monitored by an error rate distribution plot [25]. Based on the error rate information, we then selected an appropriate composition of discriminating genes with the lowest error rate. Once a signature is defined, the result of prediction strength (PS) analysis for each sample was shown. The PS values range from -1 to +1, where higher absolute values reflect stronger predictions [25]. An overview of the results for samples in different groups was then illustrated by a PS plot [25].

Classical multidimensional scaling (MDS) is performed by the standard function of the R program to provide a visual impression of how the various sample groups are related. The average linkage distance between samples is calculated by the Pearson correlation subtracted from unity to provide bounded distances in the range (0, 2), as described in our previous study [22]. The distance between two groups of samples is calculated using the average linkage measure (the mean of all pair-wise distances (linkages) between members of the two groups concerned). The standard error of the average linkage distance between two groups (the standard deviation of pairwise linkages divided by the square root of the number of linkages) is quoted when inter-group distances are compared in the text.

Immunohistochemical staining

Staining was performed on formalin-fixed, paraffin-embedded specimens using anti-ERBB2 primary antibody (DAKO, Carpinteria, CA, USA). Scoring was performed as follows. 0: undetectable staining or membrane staining in <10% of the tumor cells. 1+: faint and incomplete membrane staining in >10% of the tumor cells; 2+: weak to moderate complete membrane staining in >10% of the tumor cells; 3+: strong complete membrane staining observed in >10% of the tumor cells. ERBB2 protein expression was categorized as negative (scores 0 and 1+), or positive (scores 2+ and 3+) [29].

Authors' contributions

SJC, TYW, and HWW designed the study project. SJC and TYW collected microarray data sets and EEC materials. SJC, TYW, CYT, and TFW executed project plan and data
analysis. SJC, TYW, MDC, and HHWW carried out data interpretation and discussion. SJC wrote the manuscript. Then HHWW modified it. All authors read and approved the final manuscript.

**Additional material**

Additional file 1  
**The discrimination ability of the 678 probe sets.** Prediction power of the 678 probe sets differentiating early and late stage samples, as well as discriminating normal endometrium and tumor tissues. Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-613-S1.JPG](http://www.biomedcentral.com/content/supplementary/1471-2164-10-613-S1.JPG)

Additional file 2  
**The annotation of probed genes and cDNAs.** Complete data of analyzed arrays and clustered genes/cDNAs. Click here for file  
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-613-S2.XLS](http://www.biomedcentral.com/content/supplementary/1471-2164-10-613-S2.XLS)

**Acknowledgements**  
The authors acknowledge the efforts of IGC and expO. This work is supported by the Mackay Memorial Hospital (MMH-HB-97-05), the National Health Research Institute (NHRI-EX97-9704B), the National Science Council (NSC97-3111-B-010-004 and NSC98-3230-B-010-020-MY3), Taipei Veterans General Hospital Research Fund, VGHUST Joint research Program, Tsou’s Foundation (V98ER-003), and Yang-Ming University (a grant from Ministry of Education, Aim for the Top University Plan).

**References**

1. Yilmaz OH, Valdez R, Theisen BK, Guo W, Ferguson DO, Wu H, Morrison SJ: Pten dependence distinguishes haematopoietic stem cells from leukaemia-initiating cells. Nature 2006, 441(7092):475-482.
2. Barnhart BC, Simon MC: Metastasis and stem cell pathways. Cancer Metastasis Rev 2006, 25(2):261-271.
3. Trosko JE: Review paper: cancer stem cells and cancer: non-stem cells from adult stem cells or from reprogramming of differentiated somatic cells. Vet Pathol 2009, 46(2):176-193.
4. Reya T, Morrison SJ, Clarke MF, Weissman IL: Stem cells, cancer, and stem cell cancers. Nature 2001, 414(6859):105-111.
5. Lee J, Kodiarov Y, Kodiarov Y, Li A, Su Q, Donin NM, Pastorino S, Purov BW, Christopher N, Zhang W, Park JH, Fine HA: Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. Cancer Cell 2006, 9(5):391-403.
6. Yang ZJ, Ellis T, Markant SL, Read TA, Kessler JD, Bourboulia D, Monahan AP, Rowitch DH, Wainwright BJ, Wechsler-Reya Rj: Medulloblastoma can be initiated by deletions of Patched in lineage-restricted progenitors or stem cells. Cancer Cell 2008, 14(2):135-145.
7. Schaller U, Heine VM, Mao J, Kho AT, Dillon AK, Han YG, Huillerd E, Sun T, Ligon AH, Qian Y, Ma Q, Alvarez-Buylla A, McMahon AP, Rowitch DH, Ligon KL: Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. Cancer Cell 2008, 14(2):123-134.
8. Perez-Caro M, Cabalada C, Gonzalez-Herrero I, Vicente-Duenas C, Bermejo-Rodriguez C, Sanchez-Beato M, Orfao A, Pintado B, Flores T, Sanchez-Martín M, Jimenez R, Piris MA, Sanchez-Garcia I: Cancer induction by restriction of oncogene expression to the stem cell compartment. Embry J 2009, 28(1):8-20.
9. Hahn WC, Counter C, Lundberg AS, Beijersbergen RL, Brooks MW, Weinberg RA: Creation of human tumour cells with defined genetic elements. Nature 1999, 400(6743):464-468.
10. Kho AT, Zhao Q, Cai Z, Butte AJ, Kim JY, Pomeroay SL, Rowitch DH, Kohane I: Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. Genes Dev 2004, 18(6):629-640.
11. Hu M, Shvidasani RA: Overlapping gene expression in fetal mouse intestine development and human colorectal cancer. Cancer Res 2005, 65(19):8715-8722.
12. Kaiser S, Park YK, Franklin JL, Halberg RB, Yu M, Josen WJ, Freudenberg J, Chen X, Haigis K, Jeggel AG, Kong S, Achtek B, Xu H, Reichling T, Aghar M, Boivin GP, Roberts RB, Bissahayco AC, Gonzales F, Bloom GC, Eschrich S, Carter SL, Aronow JE, Kleimeyer J, Kleimeyer M, Ramaswamy V, Settle SH, Boone B, Levy S, Graff JM, et al: Transcriptional recapitulation and subversion of embryonic development by mouse colon tumor models and human colon cancer. Genome Biol 2007, 8(7):r131.
13. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA: An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 2008, 40(5):499-507.
14. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, Campbell LL, Polyak K, Brinken C, Yang J, Weinberg RA: The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 2008, 133(4):704-715.
15. Morel AP, Lievre M, Thomas C, Hinkel G, Ansieau S, Puisieux A: Generation of breast cancer stem cells through epithelial-mesenchymal transition. PLoS ONE 2008, 3(8):e3088.
16. Liu FS: Molecular carcinogenesis of endometrial cancer. Tai- wan J Obstet Gynecol 2007, 46(1):26-32.
17. Sorosky Jl: Endometrial cancer. Obstet Gynecol 2008, 111(2 Pt 1):435-447.
18. Barker N, Ridgway RA, van Es JH, Watering M van de, Bogelh H, Born M van den, Danenberg E, Clarke AR, Sansom OJ, Clevers H: Crypt stem cells as the cells-of-origin of intestinal cancer. Nature 2009, 457(7229):608-611.
19. Liu Z, Gibson P, Currie DS, Tong Y, Richardson RJ, Bayazitov IT, Popicotet H, Zachkuteko S, Ellison DW, Gilbertson RJ: Ptenomin I marks intestinal stem cells that are susceptible to neoplasic transformation. Nature 2009, 457(7229):603-607.
20. Flier LG van der, Haegebarth A, Stange DE, Watering M van de, Clevers H: OLFM4 is a robust marker for stem cells in human stem cells and marks a subset of colorectal cancer cells. Gastroenterology 2009, 137(1):15-17.
21. Wang HW, Trotter MW, Lagos D, Bourboula D, Henderson S, Makinen T, Elliman S, Flanagan AM, Alitalo K, Boshoff C: Kaposi sarcoma herpesvirus-induced cellular reprogramming contributes to the lymphatic endothelial gene expression in Kaposi sarcoma. Nature 2004, 437(7059):673-679.
22. Huang TS, Hishey JH, Wu YH, Jen CH, Tsang YH, Chou SH, Partanen J, Anderson H, Jaatinen T, Yu YH, Wang HW: Functional network reconstruction reveals somatic stemness genetic maps and dedifferentiation-like transcriptional reprogramming induced by GATA2. Stem Cells 2008, 26(5):1186-1201.
23. Golub TR, Slonim DK, Tamayo P,uard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caliguri MA, Blobfeld CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286(5439):531-537.
24. Ramaswamy S, Ross KN, Lander ES, Golub TR: A molecular signature of metastasis in primary solid tumors. Nat Genet 2003, 33(4):49-54.
25. Jen CH, Yang TP, Tung CY, Lin CH, Hsu MT, Wang HW: Signature Evaluation Tool (SET): a java-based tool to evaluate and visualize the sample discrimination abilities of gene expression signatures. BMC Bioinformatics 2008, 9(1):58.
26. Shabo I, Stal O, Olsson H, Dvore S, Swaroop I, Bream, cancer expression of CD163, a macrophage scavenger receptor, is related to early distant recurrence and reduced patient survival. Int J Cancer 2008, 123(4):780-786.
27. Kohomara Y, Ohnishi K, Kuratsu J, Takeya M: Possible involvement of the ME2 anti-inflammatory macrophage phenotype in growth of human gliomas. J Pathol 2008, 215(1):15-24.

28. Uharcsek P: Prognostic factors in endometrial carcinoma. J Obstet Gynaecol Res 2008, 34(5):776-783.

29. Grushko TA, Filici I, Mundt AJ, Ridderstraal K, Olopade OI, Flemming GF: An exploratory analysis of HER-2 amplification and overexpression in advanced endometrial carcinoma: a Gynecologic Oncology Group study. Gynecol Oncol 2008, 108(1):3-9.

30. Strojan P, Budina M, Smid L, Svetic B, Vrhovec I, Kosi J, Skrj: Prognostic significance of cysteine proteinase cathepsins B and L and their endogenous inhibitors stefins A and B in patients with squamous cell carcinoma of the head and neck. Clin Cancer Res 2000, 6(3):1052-1062.

31. Byers RJ, Sakiha E, Joseph P, Glennie C, Hoyland JA, Menasce LP, Radford RJ, Illidge T: Clinical quantitation of immune signature in follicular lymphoma by RT-PCR-based gene expression profiling. Cancer Res 2000, 60(10):673-689.

32. Salvesen HB, Akslen LA: Molecular pathogenesis and prognostic factors in endometrial carcinoma. ApMIS 2002, 110(10):573-689.

33. An C, Choi IS, Yao JC, Worah S, Xie K, Mansfield PF, Ajani JA, Rashid A, Uharcek P, Devilee P: Prognostic significance of CpG island methylator phenotype and microsatellite instability in gastric carcinoma. Clin Cancer Res 2005, 11(2 Pt 1):656-663.

34. Toyota M, Kopecky Kj, Toyota MO, Jair KW, Willman CL, Issa JP: Methylation profiling in acute myeloid leukemia. Blood 2001, 97(19):2823-2829.

35. Bieche I, Lerebours F, Touzi S, Espie M, Marty M, Lidereau R: Molecular profiling of inflammatory breast cancer: identification of a poor-prognosis gene expression signature. Clin Cancer Res 2004, 10(20):7679-7690.

36. Stubbens RW: Cell biology of cancer: BRCA1 and sister chromatid pairing reactions? Cell Cycle 2008, 7(4):449-452.

37. Juric D, Sale S, Hromas RA, Yu R, Wang Y, Duran GE, Tishbiri SH, Einhorn LH, Slikov BI: Gene expression profiling differentiates germ cell tumors from other cancers and defines subtype-specific signatures. Proc Natl Acad Sci USA 2003, 102(49):17763-17768.

38. Kho DH, Abe J, Lee JH, Cho HJ, Cho SH, Seo YW, Ahn KY, Chung IJ, Kim KK: KIT/TEK in breast cancer: identification of a poor-prognosis gene expression signature. Cancer Res 2000, 60(10):3676-3679.

39. Birnie R, Bryce SD, Roome C, Dussupt V, Droop A, Lang SH, Berry PA, Hyde CF, Lewis JL, Stower MJ, Maitland NJ, Collins AT: Gene expression profiling of primary breast tumours maintained in distant metastases. Proc Natl Acad Sci USA 2003, 100(26):15901-15905.

40. Podsypanina K, Du YC, Jechlinger M, Beverly LJ, Hambardzumyan D, Varmus H: Seedling and propagation of untransformed mouse mammary carcinoma cells in the lung. Science 2008, 312(5787):1841-1844.

41. Husemann Y, Geigl JB, Schubert F, Musiani P, Meyer M, Burghart E, Forni G, Eils R, Felten T, Rietmuller G, Klein CA: Systemic spread of breast cancer: early step in breast cancer. Cancer Cell 2008, 13(1):58-68.

42. Chang SJ, Huang TS, Wang KL, Wang YC, Yang YC, Chang MD, Wu YH, Wang HW: Genetic network analysis of human CD34+ hematopoietic stem/progenitor cells. Taiwanese Journal of Obstet Gynecol Res 2008, 47(4):422-430.

43. Yang TP, Chang TY, Lin CH, Hsu MT, Wang HW: ArrayFusion: a web application for multi-dimensional analysis of CGH, SNP and microarray data. Bioinformatics 2008, 24(21):2697-2698.

44. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Nil L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, et al.: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004, 32(1):258-261.

45. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003, 4(5):P3.

46. Hoffman AE, Zheng T, Ba Y, Zhu Y: The circadian gene NPAS2, a putative tumor suppressor, is involved in DNA damage response. Mol Cancer Res 2008, 6(9):1461-1468.

47. Devesane AM, Akunowicz JD, Reveles XT, Patel BB, Sara EA, Gorklick RG, Naylor SL, Leach RJ, Hansen MF: PHC3, a component of the hPRC-H complex, associates with EZF6 during G0 and is lost in osteosarcoma tumors. Oncogene 2007, 26(12):1714-1722.

48. Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, Chen CL, Yu CJ, Lee YC, Chen HS, Su TJ, Jiang CC, Li HN, Hong QS, Hu HY, Chen CC, Chen WJ, Liu CC, Chan WK, Li KC, Chen JJ, Yang PC: A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007, 356(1):1-20.

49. Weigelt B, Glas AM, Wessells LF, Witteveen AT, Peterse JL, van't Veer LJ: Gene expression profiles of primary breast tumours maintained in distant metastases. Proc Natl Acad Sci USA 2003, 100(26):15901-15905.