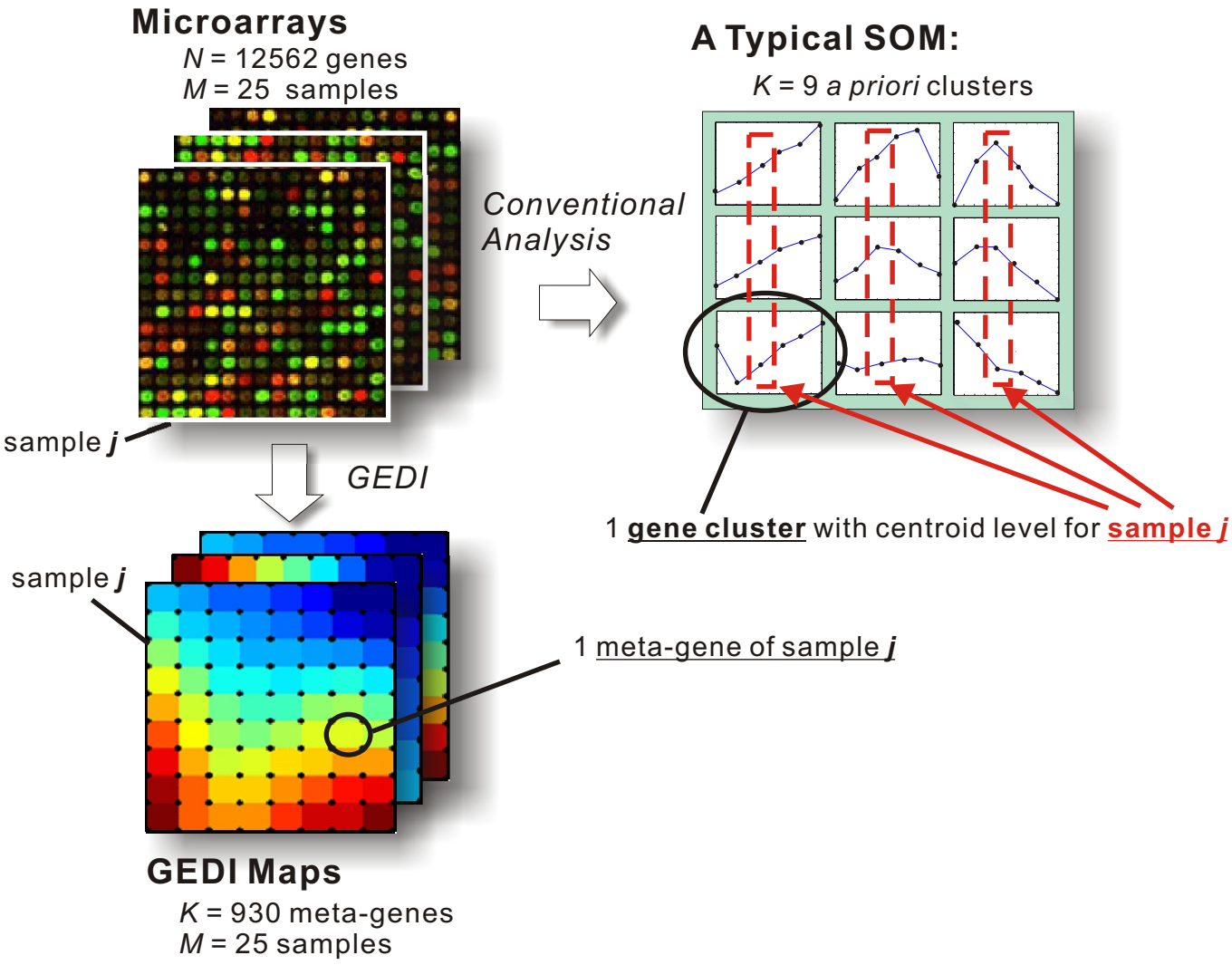

Figure S1 - Principle of GEDI analysis
The program GEDI maps each molecular profile of a given sample measurement (e.g., microarray) into a "GEDI map" (mosaic) with characteristic color patterns generated by the ordering of the meta-genes represented by the individual "tiles" of the mosaic. Each tile is a formal SOM "mini-cluster". Colors represent value of the centroid expression levels. Note that biologically relevant "clusters" of either genes or samples are recognized by visual inspection of the GEDI maps based on higher-order "meta-patterns" and are not pre-defined by the SOM grid structure as in conventional SOM where the information of the entire stack of M microarray is mapped into one grid structure with *K a priori* clusters.

GEDI Analysis Parameters

The parameters used in GEDI for creating the GEDI maps shown in the main text of the paper are as follows (Table S1):

For detailed instructions on GEDI, see manual at:
http://web1.tch.harvard.edu/research/ingber/GEDI/gedihome.htm

GEDI parameter	Value
x_grid	31
y_grid	30
som_train_num1	20
som_train_num2	80
neighborhood	8.0
neighborhood2	1.0
beta (learning factor)	0.5
beta2 (learning actor)	0.05
conscience	3.0
conscience2	3.0
nbh_block	4
nbh_block2	2
init_method	Linear Initialization
random_seed	1
distance_metrics	Euclidean distance
clip	0.0010

Table S1. GEDI parameters used for the GEDI maps shown in the main text.

GEDI Maps with varying SOM parameters

To evaluate the stability of the GEDI maps, we performed the GEDI analysis of the same data set with varying sets of SOM parameters. Below are three examples in which neighborhood, learning factor or SOM grid size were varied. The results show that when the SOM was run to convergence, GEDI preserved the distinctness of patterns with the same inter-sample relationships, albeit the individual patterns look different in the different runs.

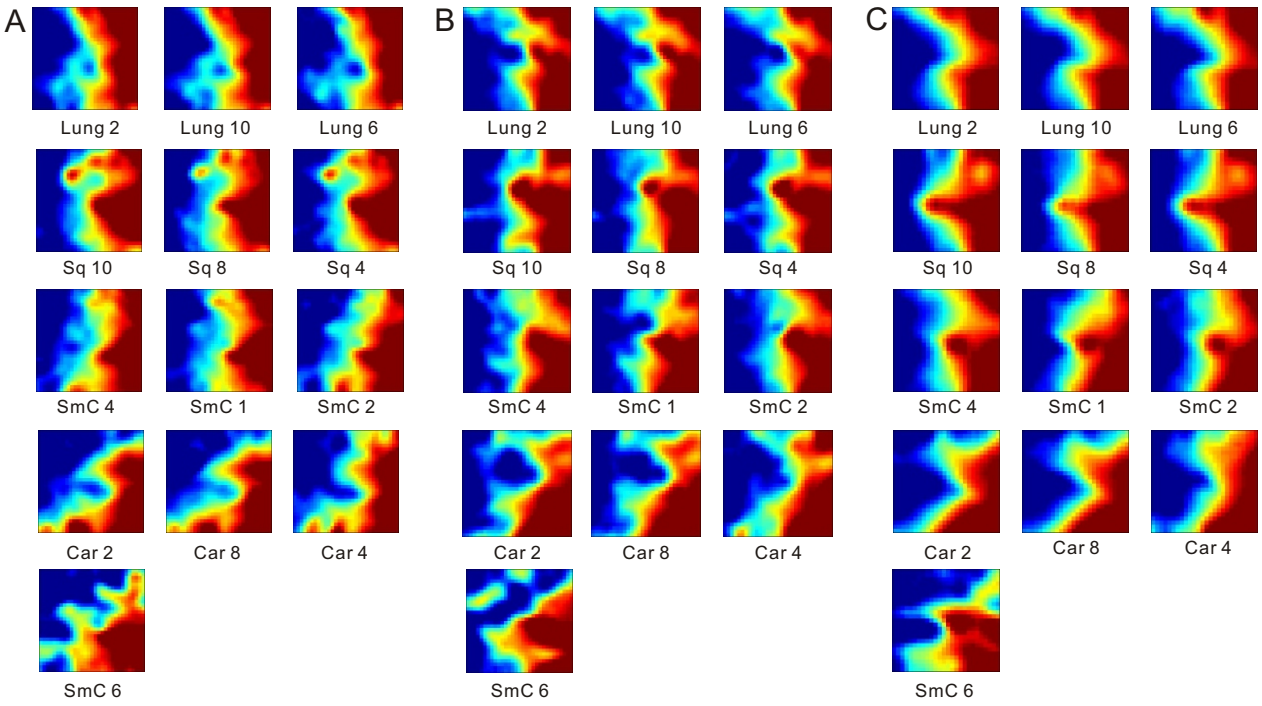

Figure S2. GEDI Maps with different SOM parameters
A. Change neighborhood parameter (n=4.0)
B. Change learning factor parameter (b=0.7)
C. Change grid parameters (x=26, y=25)

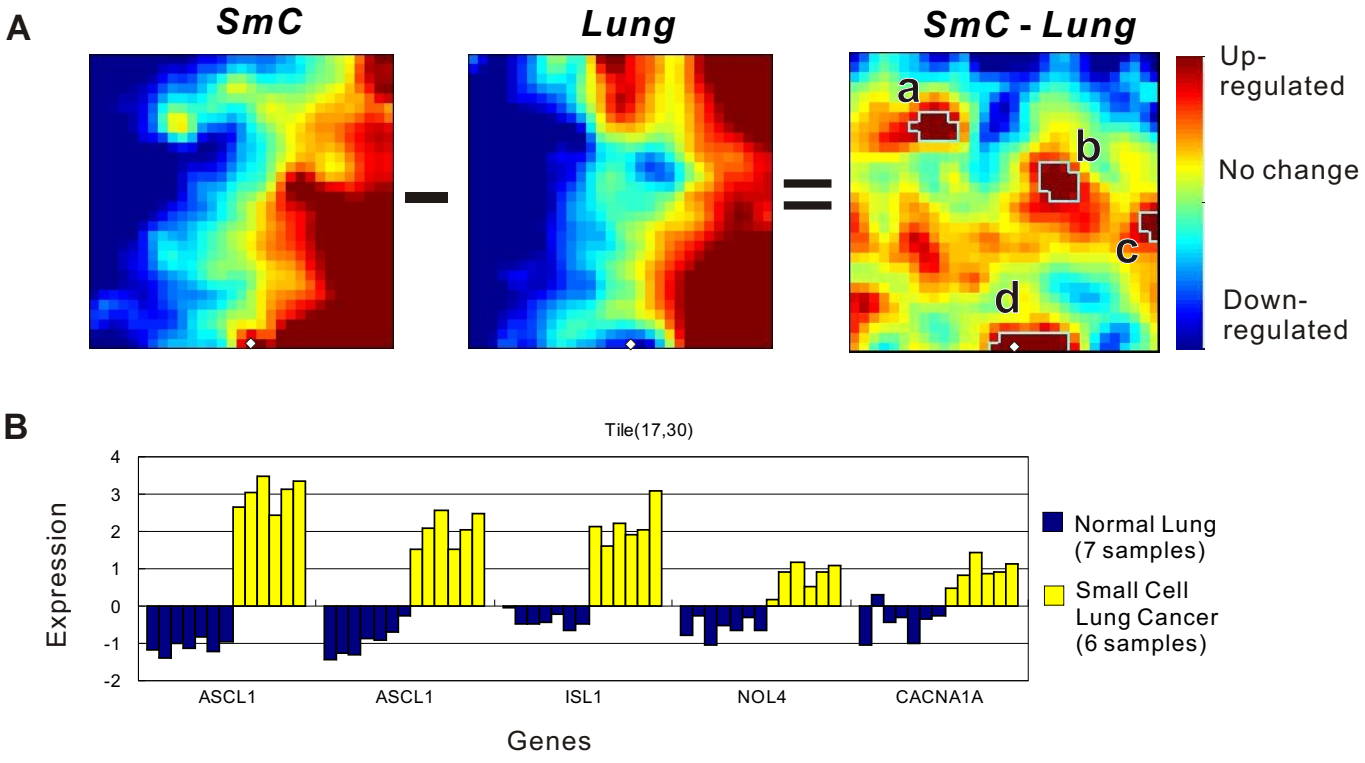

Figure S3 - GEDI average maps and difference maps

A. GEDI average maps and difference maps. The colors of the miniclusters are based on average of the meta-gene (centroid) values across all 6 *SmC* and 7 *Lung* maps. The *SmC-Lung* map was obtained by element-wise subtraction of the average maps on the left of the equal sign. The circled regions a-d show four "Gene islands" containing the 5% most up-regulated genes in SmC compared to normal lung. Color bar represents centroid values after subtraction.

B. Retrieval of gene-specific information in GEDI. Representation of expression levels of genes in map unit (17, 30) in a form similar to what is displayed by the GEDI program when a map unit (minicluster) is clicked. The height of the bars represent the gene expression values (here, as log2 transformed values). Each bar represents one individual patient sample. The similarity in gene expression in this minicluster between the samples is manifest in that all 5 genes it contains were up-regulated in small cell lung cancers (yellow bars) compared to normal lung tissues (blue bars).

Functional enrichment analysis of GEDI "Gene islands"

To evaluate the biological relevance of GEDI "gene islands", we used the online tool EASE (http://david.niaid.nih.gov/david/ease.htm) [1] to systematically investigate the functions of the genes in the up-regulated regions on GEDI maps. In our analysis, EASE identifies Gene Ontology (GO) annotations (biological processes, molecular functions and cellular components) that describe a statistically significant number of genes in the GEDI gene island with respect to the number of genes described by the term or phrase in the population of genes on the whole gene chip. An EASE score, an adjusted Fisher exact probability, is used to report statistical measures of over-representation.

We obtained from GEDI four gene lists from the 4 most up-regulated (*SmC* vs *Normal Lung*) regions and used EASE [1] to find out their biological themes. The Excel spreadsheets in Supplementary information file 2 show the EASE biological themes of these four up-regulated gene islands in small cell lung cancers. Growth related gene functions (cell proliferation, cell cycle, DNA replication, etc.) are over-represented in region *a*, *b* and *c* whereas in region *d*, a totally different category of neural tissue genes (synaptic vesicle, neuromuscular physiological process, etc.) are enriched. This is consistent with the neuroendocrine features of small cell lung cancer.

Excel Spreadsheets (Supplementary information file 2).
The columns "Xcoord" and "Ycoord" refer to the GEDI map position as shown in Fig 4 and S3. Each sheet contains a gene list of one of the four "gene island" as indicated on the tabs at the bottom.

1. Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: Identifying biological themes within lists of genes with EASE. Genome Biol 2003, 4(10):R70

Comparison of Sq2 with other Sq samples

Visual inspection of GEDI maps readily confirms the notion that Sq2 displays significant feature similarity to the SmC samples. Specifically, in the marked rectangular areas (column 11~14, row 7~10), the gene expression activities of Sq2 differs from that in other Sq samples (overexpression), but is similar to that of the SmC sample (suppression) (Figure S4). The miniclusters in this area contain several keratin genes (KRT6A, KRT14, KRT15, KRT16), biomarkers of squamous lung cancer, that are highly expressed (red in Fig S4A) in all Sq samples except in Sq2. This points to some problem with classifying Sq2 as squamous lung cancer even if conventional hierarchical cluster analysis using aggregate information of expression profiles places Sq2 under the cluster of squamous lung cancer.

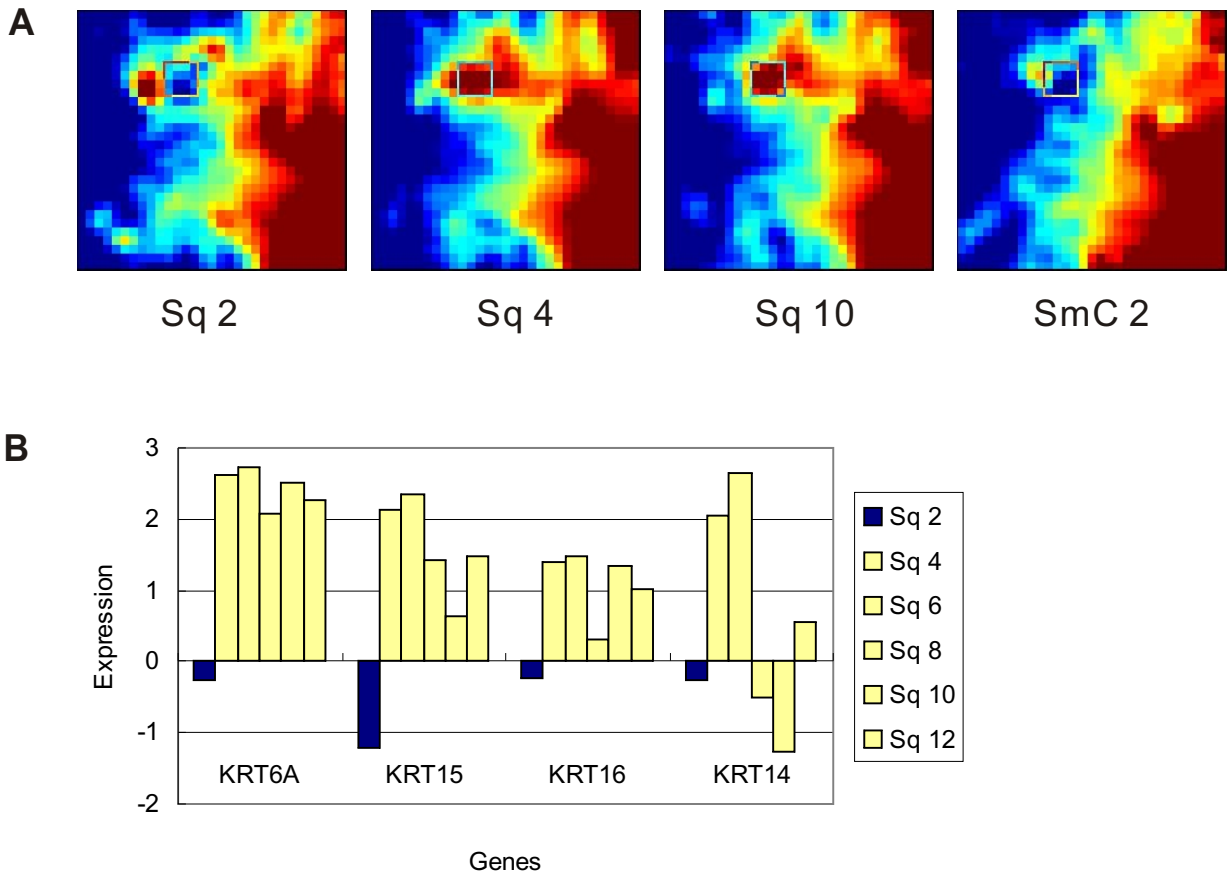

Figure S4. Comparison of Sq2 with other Sq samples

A. GEDI maps with highlighted area (rectangle of 4x4 miniclusters) as shown in the main text, but magnified to reveal a distinct difference between Sq2 and other Sq samples. Suppression of genes in this area in Sq2 recapitulates the behavior of genes in the SmC sample.

B. The relative expression values (log2-transformed values) of 4 keratin genes found in the highlighted area for all the Sq samples.