**Figure Legends**

**Figure 1 EPIG-Seq workflow and example.** A) It exemplifies the strategy employed for EPIG-Seq, where two main goals were achieved throughout the process with pre-defined parameters used. The output was the extracted significant patterns supported by genes and associated with p-value as significance. B) We obtained two transcripts from an RNAseq experiment (noted as transcript A & B) with four conditions (noted as group); four lanes of RNAseq data were involved under each condition, 16 lanes total. Firstly, we tried to detect “zero”, which represents zero count for a specific transcript at a condition. There were 12 lanes with no zero count (a), two lanes with zero count mapped to transcript A from condition 1 (b) and two lanes with zero count mapped to transcript B from condition 2. Intermediate distance measures Di, Dk, D1, D2, and D3 were also computed and finally CYd, CYobs and CY were computed.

**Figure 2** **Simulated RNA-Seq data with six patterns** Six distinguishable patterns were simulated for evaluating the algorithm in panel 1 – 6, where pattern 1-5 carry meaningful biological information and last pattern represents the noise pattern. In each panel, four groups are laid out on the x-axis and the y-axis indicates the relative log2 ratio as shown in table 2.

**Figure 3 Unsupervised analyses on gene profiles from simulated data with five patterns and one background noise.** 3A. Principal component analysis was performed on CY as the correlation distance. Each group (baseline, group 2-4) had 30 simulated lanes mimicking biological replicates were color-coded: Baseline (red), group 2 (blue), group 3 (green), and group 4 (purple). Each lane had 20000 transcripts. 3B Heatmap on simulated data of all six seeded patterns, each with 200 transcripts. Four different groups (baseline, group 2 -4 ) were laid out along the horizontal order. On the vertical scale, six patterns were aligned with each pattern of 200 transcripts. Six distinguishable pattern (A – F) were color coded by the vertical side bar. The heatmap gradient scale shows log of magnitude relative to the baseline.

**Figure 4 Parameter-space searching for the optimized default choice for end users.** 4A. Parameter optimization in step one: there were six panels with labels on the top indicating the choice of location parameter St1 at [5-10]; within each panel, the x-axis shows the CYs as the similarity measure [0.5 – 0.9]; the color code was for the two-tailed dispersion cut off at 1 – 5 %. The y-axis in each panel was the Adjusted Rand Index [0-1]. 4B. Parameter optimization in step two, x-axis shows the CYs as the similarity measure [0.5 – 0.9]; the y-axis was the Adjusted Rand Index [0-1]; the color code was for indicating the choice of location parameter St2 at [2-5].

**Figure 5 EPIG-Seq results running from the simulated dataset (aforementioned).** 5A. The thumbnails of the five distinguishable patterns extracted by EPIG-Seq from the simulated data. Four groups were color-coded: Baseline (red), group 2 (blue), group 3 (green), and group 4 (purple), 30 lanes were simulated in each group. Each panel represents a pattern, which was supported by a number of transcripts extracted from EPIG-Seq method, four groups are laid out on the x-axis and the y-axis indicates log2 ratio. 5B. The heatmap plotted on the combined extracted transcripts from all five patterns, the heatmap gradient scale shows log of magnitude relative to the baseline. Horizontally, four groups were clustered and laid out side-by-side. To the right of the heatmap, six known patterns were color-coded in a vertical bar, another color bar indicates which pattern the extracted transcript was assigned to. 5C. PCA analysis was performed on the aforementioned combined extracted transcripts using the pair-wise CYs as the covariance matrix.

**Figure 6 The application of EPIG-Seq on the SEQC toxicity data**. 6A. Thumbnails of the four significant patterns extracted by EPIG-Seq (six MOA were color-coded) from the SEQC toxicity data. Within each panel, six chemical perturbation groups are laid out on the x-axis indexed by the lanes and the y-axis indicates log2 ratio. 6B The heatpmap plotted on the combined extracted transcripts from all four patterns, the heatmap gradient scale shows log of magnitude relative to the baseline. Horizontally, six chemical perturbation groups were clustered and laid out side-by-side. The transcripts extracted from all four patterns collectively shown in gene symbol to the right of the heatmap. 6C. PCA analysis was performed on the aforementioned combined extracted transcripts using the pair-wise CYs as the covariance matrix.

Figure 7. Immunohistochemistry staining of ITGB1 in breast cancer tissue and normal tissue A) normal breast tissue. B) Cancer breast tissue.