Figure 1 EPIG-Seq workflow

Figure 2 Simulated RNA-Seq data with five patterns (pattern 1-5) carry biological meaning and pattern 6 as background noise

Figure 3 Unsupervised analyses on gene profiles from simulated data with five patterns and background noise. A. A conventional PCA analysis was performed on pair-wise CYs as the covariance matrix. B. A hierarchical clustering reveals prominent patterns in simulated data

Figure 4 Parameter-space searching for the optimized default choice for users. A. Parameter optimization in step one, there were four panels with labels on the top indicating the choice of location parameter St1 at [1-5], within each panel, the x-axis shows the CYs as the similarity measure [0.5 – 0.9], where 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]. B. Parameter optimization in step two, x-axis axis shows the CYs as the similarity measure [0.5 – 0.9] and the y-axis in each panel was the Adjusted Rand Index [0-1].

Figure 5 EPIG-Seq results running from the simulated dataset (aforementioned). A. The thumbnails of the 5 extracted simulated patterns extract by EPIG-Seq. Group1 (labeled as “Baseline”) in red, group2 in green; group3 in blue and group4 in purple. B. The hierarchical clustering reveals prominent patterns in simulated data. C. conventional PCA analysis was performed on pair-wise CYs as the covariance matrix.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| TranscriptA | 0 | 16 | 1 | 0 | 1228 | 746 | 166 | 574 | 5207 | 25693 | 1018 | 2622 | 364608 | 58221 | 210304 | 239869 |
| TranscriptB | 36 | 328 | 190 | 178 | 36 | 4 | 0 | 0 | 2581 | 7068 | 298 | 17049 | 23 | 2542 | 799 | 931 |

Table 1. Count level measurement of two transcripts (genes) in four groups

|  |  |  |  |
| --- | --- | --- | --- |
|  | EPIG | EPIG-seq | ORIGEN |
| Data type | Continuous | Count level | Count level |
| Distribution assumption | Gaussian | Poisson | NonParametric |
| Correlation measurement | Person's | CYs | Order statistics |
| Measurement or spreading | Variance | Dispersion | Residue |
| Magnitude measurement | LogRatio | Wilcoxon test | Difference Between means |
| Significant test | SignalToNoise | Variance-to-mean ratio (VMR) | mdFDR means |
| Deliverable results | Co-expression pattern | Co-expression pattern | Co-expression pattern |

Table 2. Comparison of methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | EPIG | EPIG-seq | SAMseq | Deseq |
| Data type | Continuous | Count level | Count level | Count level |
| Distribution assumption | Gaussian | Poisson | Poisson | Negative Binomial |
| Correlation measurement | Person's | Cys | N/A | N/A |
| Measurement or spreading | Variance | Dispersion | Dispersion | Dispersion |
| Magnitude measurement | LogRatio | Wilcoxon test | Wilcoxon test | FoldChange |
| Significant test | SignalToNoise | Variance-to-mean ratio (VMR) | NonParametrics | NegBinom test |
| Deliverable results | Co-expression pattern | Co-expression pattern | DEG | DEG |

Table 2 (addition??). Comparison of methods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pattern number | Number of genes | Group1 | Group2 | Group3 | Group4 |
| 1 | 200 | 1 | 1.5 | 2.5 | 4 |
| 2 | 200 | 1 | -1.5 | -2.5 | -4 |
| 3 | 200 | 1 | -1.5 | 4 | 2.5 |
| 4 | 200 | 1 | 4 | 2.5 | 1.5 |
| 5 | 200 | 1 | 4 | 4 | 4 |
| 6 | 19000 | 1 | 1 | 1 | 1 |

Table 3. Simulated RNA-seq data with 5 meaningful patterns

Need to update on table 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Seeded |  |  |  |
| Extracted | Group\_A | Group\_B | Group\_C | Group\_D | Group\_E | Group\_F |
| Group\_A | 147 | 0 | 81 | 0 | 48 | 40 |
| Group\_B | 0 | 172 | 0 | 0 | 0 | 17 |
| Group\_C | 20 | 0 | 92 | 0 | 11 | 6 |
| Group\_D | 0 | 0 | 0 | 169 | 30 | 8 |
| Group\_E | 11 | 1 | 5 | 12 | 100 | 1 |

Table 4. EPIG-seq results on simulated RNA-seq data with 5 meaningful patterns (Group A – E)