RA dataset – GSE42861

The RA dataset was obtained from GEO and processed in accordance with the methodology described in the HIRE paper. Processing steps included:

* Removing samples GSM1051535 and GSM1051691 (due to no smoking history)
* Removing CpG sites with a high methylation mean across samples (>0.8) and low methylation mean (<0.2)
* Batch affect adjustment using COMBAT – the 10,000 most variable sites were kept
* Edit sample information:
  + RA status: 1 for affected, 0 for normal
  + Gender: 1 for male, 0 for female
  + Smoking history:
    - Never: 1 for yes, 0 for no
    - Ex: 1 for yes, 0 for no
    - Current: 1 for yes, 0 for no
    - Occasional: 1 for yes, 0 for no

Matching cell types across the two outputs

The estimated cell-type specific methylation profiles generated by HIRE (mu\_t) and TCA (mean value for each cpg) were compared using a linear model. The list of CpGs taken forward for this analysis were the top 50 most discriminative sites as identified by GLINT.

As a proof of principle, cell type 1 from HIRE and TCA were regressed against their full datasets. As expected, cell type 1 showed the strongest association with cell type 1 in both cases.

*Regressing HIRE cell type 1 against the HIRE dataset*

summary(fit\_hire)

Call:

lm(formula = test\_hire ~ ., data = data.frame(hireoutput))

Residuals:

Min 1Q Median 3Q Max

-9.549e-17 -1.639e-17 9.760e-19 1.698e-17 2.759e-16

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.884e-16 9.930e-17 -1.897e+00 0.0645 .

celltype1 1.000e+00 1.311e-16 7.626e+15 <2e-16 \*\*\*

celltype2 -2.471e-16 2.632e-16 -9.390e-01 0.3531

celltype3 1.475e-16 8.992e-17 1.641e+00 0.1082

celltype4 -5.375e-21 2.772e-21 -1.939e+00 0.0591 .

celltype5 -1.314e-16 2.062e-16 -6.370e-01 0.5274

celltype6 3.913e-17 8.527e-17 4.590e-01 0.6486

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.273e-17 on 43 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 4.532e+31 on 6 and 43 DF, p-value: < 2.2e-16

**Warning message:**

**In summary.lm(fit\_hire) :**

**essentially perfect fit: summary may be unreliable**

*Regressing TCA cell type 1 against the TCA dataset*

> summary(fit\_tca)

Call:

lm(formula = test\_tca ~ ., data = data.frame(tcaoutput))

Residuals:

Min 1Q Median 3Q Max

-4.604e-16 -1.094e-16 -2.820e-17 1.053e-16 1.088e-15

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.024e-16 4.054e-16 1.239e+00 0.2219

celltype1 1.000e+00 4.482e-16 2.231e+15 <2e-16 \*\*\*

celltype2 -2.831e-15 1.375e-15 -2.059e+00 0.0456 \*

celltype3 1.204e-16 3.122e-16 3.860e-01 0.7017

celltype4 1.139e-15 5.244e-16 2.173e+00 0.0354 \*

celltype5 2.690e-16 4.047e-16 6.650e-01 0.5099

celltype6 6.090e-16 5.120e-16 1.190e+00 0.2407

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471e-16 on 43 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 5.502e+32 on 6 and 43 DF, p-value: < 2.2e-16

**Warning message:**

**In summary.lm(fit\_tca) : essentially perfect fit: summary may be unreliable**

Each individual cell type was then regressed against the alternative dataset. The results indicated no real cell-type matches across the two datasets. A summary of the results can be seen in the matrix below. The result to the left of each “|” is the significance code generated from TCA, and to the right, HIRE.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | HIRE |  |  |  |  |
|  |  | CT1 | CT2 | CT3 | CT4 | CT5 | C6 |
|  | CT1 | \*\*|. | | | |\* | | | | | | |
|  | CT2 | .| | \*\*|\*\* | \*\*| | | | \*| | \*\*|\*\*\* |
|  | CT3 | .| | \*\*\*| | \*\*\*|\*\*\* | | | \*\*\*|\*\* | \*\*\*|\*\*\* |
| TCA | CT4 | | | | | | | | | | | |\*\*\* |
|  | CT5 | |\* | | | \* | | | | | |\*\*\* |
|  | CT6 | \*\* | | | |\* | | | | | |\* |

The highest number of highly significant matches are amongst HIRE cell types and TCA cell type 3. It is possible that TCA cell type 3 is actually an aggregation of signals from the matched TCA cell types (3, 5 and 6).

The results here suggest that HIRE and TCA are identifying very different cell-type signatures.

Matching cell types against a blood reference dataset

The two outputs were regressed individually against the reference dataset to identify the cell types present in the mixtures. The CpG sites used for this analysis were those present in the ranked list that were covered by the reference dataset.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Gran | CD4 | CD8 | CD19 | CD14 | CD56 | Neu | Eos |
| HIRE1 |  | . |  |  |  |  |  |  |
| HIRE2 |  | . | \* |  |  |  |  |  |
| HIRE3 |  |  |  |  |  |  |  |  |
| HIRE4 |  |  |  |  | \*\*\* | \*\* |  | . |
| HIRE5 |  |  |  |  |  |  |  |  |
| HIRE6 |  |  |  |  |  |  |  |  |
| TCA1 |  |  |  |  |  |  |  |  |
| TCA2 |  |  |  |  |  |  |  |  |
| TCA3 |  |  |  |  |  |  |  |  |
| TCA4 |  |  |  |  |  |  |  |  |
| TCA5 |  |  |  |  |  |  |  |  |
| TCA6 |  |  |  |  |  |  |  |  |

There were very few matches between the cell types generated by HIRE and TCA with the reference dataset. The only highly significant match was HIRE cell type 4, identifying with CD14 cells. All of the matches (ranging from low to high significance) were HIRE cell-types suggesting HIRE was better at separating out the aggregated signals into real cell types for this particular dataset. A larger number of datasets should be tested to properly compare the performance of the two tools.