

RNAseq Downstream Analysis

Machine learning methods (including clustering, dimensionality reduction, classification and regression modeling, resampling techniques, etc.), ANOVA modeling, and empirical Bayes analysis. 12 hour minimum (\$876 internal, \$1116 external) per project.

Unsupervised Analysis

Unsupervised methods provide exploratory data analysis useful for getting a big picture view: can provide valuable QC information and can help to both assess expected trends and identify unexpected patterns in your data.

- **Deliverables:**
 - Plots in png and pdf format
 - Results from any additional algorithms applied may be provided in tab-delimited or excel formatted tables as appropriate
- **Tools Used:**
 - **Hierarchical Clustering:** both of genes and samples.
 - **Principal Components Analysis:** PCA biplot of data after centering both on the gene and sample axes (and optionally scaling of gene axis if desired).
 - **Other methods:** (e.g., k-means clustering, self-organized maps, multidimensional scaling, etc.) available if desired

Empirical Bayes Differential Expression Analysis

RNAseq experiments yield simultaneous measurements of many intrinsically similar variables (gene expression levels) but with often limited sample sizes. Empirical Bayes methods provide a statistical approach designed just for such situations which "borrow strength" across genes to increase statistical power and decrease false discovery.

- **Deliverables:**
 - Tables of model parameters, p-values, and FDR q-values (in tab-delimited and excel format)
 - Boxplots (stratified by sample group) and pairs plots of top genes provided in png and pdf format
- **Tools Used:**
 - **Limma:** applies empirical Bayes methods in the construction of linear models (e.g, t-tests, ANOVA) for a large variety of experimental designs. Originally designed for microarray data analysis, Limma's developers have substantially extended its functionality into the realm of RNAseq as well.

Supervised Analysis

Many methods available for classification and regression as appropriate to your analysis. Model performance may be assessed using standard metrics evaluated under cross-validation or using independent test sets if available. Analysis will be conducted using R and/or Python scripts.

- **Deliverables:**
 - Tables of results (in tab-delimited and excel formats)
 - plots in png and pdf format
 - R and/or Python source files
 - binary, JSON, or XML representations of R or Python objects can be made available if desired
 - further reports in the form of slides or text documents may be provided in standard formats (pdf, doc, ppt) if desired
- **Methods Available:**
 - Diagonal linear discriminant analysis (**DLDA**, a form of linear naive Bayes classification)
 - Linear and quadratic **discriminant analysis**
 - **Logistic regression** including **L1/lasso** and/or **L2/ridge** regularization if desired
 - Partial least squares (**PLS**) discriminant analysis and regression
 - *k*-nearest neighbors (**KNN**)
 - Support vector machines (**SVM**)
 - Decision tree ensembles (**Random Forests** or **AdaBoost**).
 - Other methods are available on request.