PVCA Documentation

June 28, 2011

Description A pipeline to analyze data sets for batch effects using the Principal Variance Component Analysis algorithm.

Author Adam B. Norberg (Institute for Systems Biology), Sheila Reynolds (Institute for Systems Biology); anorberg@systemsbiology.org

1 Summary

Principal Variance Component Analysis (PVCA) is an algorithm to identify the contributors to the variance of a family of values, equivalent to how Principal Component Analysis identifies linear contributions to the value itself. This implementation is designed to look for batch effects- influences on the data that can be traced to *how it was collected* as opposed to attributes inherent to that which is being measured. Once batch effects have been identified, they can be controlled for. This pipeline only identifies whether or not such controls are necessary.

This algorithm is implemented in R, due to its strong statistics library and intuitive support for matrix operations. However, R is extremely bad at handling divergent and awkward file formats, a fact pointed out in R's documentation itself. The beginning of section 7 of R's guide recommends using a different programming language to convert data to a format easily processed by R, so a preprocessor that handles the divergent formats we expect to need to handle has been written in Java, and forms the first step of the pipeline. Producing the output report has also been isolated into its own module.

2 Parameters

Input File A tab-delimited gene expression or methylation data file, with exactly one or two header rows or columns. Features must be along the rows, with TCGA samples along the columns. The top column header must be the TCGA sample barcode. Methylation data has a great deal of special code handling it and must be considered highly fragile. Due to a lack of formal data standards, parsing is on a best-effort basis.

Keep Rate Either a floating-point value between 0 and 1, representing the proportion of features to maintain in the data set, or an integer 2 or greater, representing the exact number of features to maintain. The specified fraction or number of features is selected based on which features show the greatest median absolute deviation from the mean.

3 Output Files

pvca.html The primary report generated by Nozzle. Provides a summary of results and a bar chart displaying what, if any, batch effects were found.

pvca.png The bar chart used by the description.

4 Platform Dependencies

Java GenePattern must use Java 1.6 or later.

R This is being built and tested against R 2.12. However, other versions are more likely than not to work.