# pvcaCore Documentation

June 29, 2011

**Description** An R implementation of the Principal Variance Component Analysis algorithm. Requires extremely specific input format.

**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 module implements the specific step of running the algorithm. pvcaPreprocessor, in Java, is a required preprocessing step, since R is fairly slow at the parsing and data rearrangement required to support a variety of file formats with the same program.

## 2 Parameters

**Input File** A gene expression or methylation dataset already processed by pvcaPreprocessor. None of the “standard” file formats will work here; this tool requires the specific format output by the preprocessor.

# 3 Output Files

**stdout** A brief report on the factors being checked for batch effects.

**stats.RData** A compressed, R-formatted data file encapsulating all the calculated data.

**pvca.png** The bar chart representing the observed batch effects.

# 4 Platform Dependencies

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