ImputeMissingValuesKNN Documentation

Description: Imputes missing data
Author: Joshua Gould, gp-help@broad.mit.edu

Summary:
The following description is from Hastie et. al (1).

For each gene with missing values, we find the *k* nearest neighbors using a Euclidean metric, confined to the columns for which that gene is NOT missing. Each candidate neighbor might be missing some of the coordinates used to calculate the distance. In this case we average the distance from the non-missing coordinates. Having found the k nearest neighbors for a gene, we impute the missing elements by averaging those (non-missing) elements of its neighbors. This can fail if ALL the neighbors are missing in a particular element. In this case we use the overall column mean for that block of genes.

References:

1. Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. Impute 1.0.2 R package, http://bioconductor.org.
2. Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D., Imputing Missing Data for Gene Expression Arrays, Stanford University Statistics Department Technical report (1999), http://www-stat.stanford.edu/~hastie/Papers/missing.pdf
3. Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman, Missing value estimation methods for DNA microarrays BIOINFORMATICS Vol. 17 no. 6, 2001 Pages 520-525

Parameters:

Name	Description
data.filename	Data file (missing values are stored as NA) - .gct
k	Number of neighbors to be used in the imputation
rowmax	The maximum percent missing data allowed in any row
colmax	The maximum percent missing data allowed in any column
output.file	The name of the output file - .gct

Platform dependencies:

Module type:	Missing Value Imputation
CPU type:	any
OS:	any
Language:	R