**Variable Key**

*E\_log10\_QN* ∷ original *G* × *N* gene expression matrix with *G* rows of genes and *N* columns of samples

*names* ∷ list of names for the *G* genes corresponding to rows in *E\_log10\_QN* (a.k.a. the set *g* = {*g*1,…,*gG*})

*groups* ∷ binary vector representing class assignments; 1 = class 1, 0 = class 2

\**\_gs\_defs* ∷ matrix containing groups of genes belonging to each gene set *m* = {1,…,*M*} for a particular collection (e.g. \* = *biocarta*)

*gs\_struct* ∷ structure created using the function *gs\_match\_id* that contains the following information for mapping gene expression from a dataset to *a priori* defined gene sets

*gs* ∷ list of names for the *M* gene sets in the collection (i.e. *gs* = {*gs*1,…,*gsM*})

*gs\_id* ∷ if gene sets have numerical IDs corresponding to names (e.g. GO IDs and terms), they are listed in this array

*G\_gs\_total* ∷ the total number of distinct genes found in the original gene set definitions matrix

*G\_gs­\_matched* ∷ the total number of distinct genes from the expression dataset mapped onto gene sets in the collection

*g\_gs* ∷ contains subsets of gene names for each gene set in the collection, where each row contains the *Gm* genes belonging to gene set *m*; each subset of genes can be denoted as *gm*

*g\_gs\_idx* ∷ same as *g\_gs*, but gene names are replaced with the specific row indexes in the expression matrix *X*

*g\_gs\_match\_rate* ∷ for each gene set *m*, lists (i) the number of dataset genes mapped to the set *Gm*, (ii) the number of genes originally listed for the set in the definitions file *Gm,0*, and (iii) the fraction of mapped genes to total genes for the gene set

*X* ∷ compressed expression matrix containing only unique genes; for duplicate gene names (i.e. those that appear in multiple rows of the original expression matrix), expression values for each sample are taken as the maximum expression among all duplicates in that sample

*g* ∷ list of unique gene names corresponding to rows in *X*