MASSI: a MicroArray Sample Sex Identifier

Sam Buckberry

September 30, 2013

1 The Problem

Given that the sex of many species is an easily observable and usually unambiguous classification, it is surprising the number of microarray datasets in public repositories that lack the associated sample sex information. Sex-biased gene expression in normal and pathological tissues is a well recognised for both sex chromosome and autosomal genes. Sex biases also exist in the prevalence and severity of many common human diseases, such as cardiovascular disease and some cancers. As sex is a potential influencing factor of both pathological and non-pathological phenotypes, gene expression analyses that do not account for sex-specific effects could fail to identify a significant proportion of genes that contribute the condition under investigation. Therefore, the absence of sample sex information restricts the reuse of gene expression datasets where the researcher intends to factor the effect of sex in reanalysis or reinterpretation, or when intending to include such datasets in larger gene expression meta-analyses.

This is why we developed MASSI, an R package for predicting the sex of samples in microarray datasets. The MASSI package allows researchers to expand their analyses to retrospectively incorporate sex as a variable, generate or confirm sex information associated with publicly available datasets, or to accurately predict the sex of samples missing sex information.

2 Importing data to use in the MASSI analysis

The MASSI analysis begins by importing a standard gene expression matrix of normalised and log2 transformed probe values. The gene expression matrix must be as a data.frame object and have the sample identifiers as the column names and the probe identifiers as the row names. The Y chromosome probe identifiers must also be as a data.frame object with the identifiers as row.names.

To load the included test MASSI gene expression matrix:

```
> library(MASSI)
> data(MASSI.test.dataset)
```

The included gene expression matrix is composed of 60 samples and 1026 probes as a data.frame object.

To load the test probe list corresponding to the included data:

```
> data(MASSI.test.probes)
```

The Included list of Y chromosome probes contains 56 probe identifiers as row.names as a data.frame object.

If expression data is stored in the ExpressionSet class, you can load the expression data into the format required for the MASSI analyis using a command such as:

```
> norm.log.values <- exprs(object=ExpressionSet)
```

3 Extracting the Y chromosome probe data

The first step of the MASSI analysis involves extracting the expression values for probes that correspond to Y chromosome genes. When the expression values for Y chromosome probes are extracted, the expression variance for each probe across all samples is calculated. This allows the identification of low variance probes, which are unlikely to be informative in sex classification. The user has the option of selecting a probe variation threshold, so only the most informative probes are used in the classification process. Deciding on a probe variation threshold can be informed by inspecting the automatically generated probe variation plot (Figure 1). In our experience, using the most variable 25-50% of probes (typically 10-40 probes, depending on platform) produces good results.

To extract data corresponsing to Y chromosome probes from the test dataset and look at a probe variation plot:

```
> massi.y(exprs=MASSI.test.dataset, y.probes=MASSI.test.probes)
```

The plot pictured in Figure 1 should be output to the R graphics device, and also as a saved file "massi.yploy.pdf" in the current working directory.

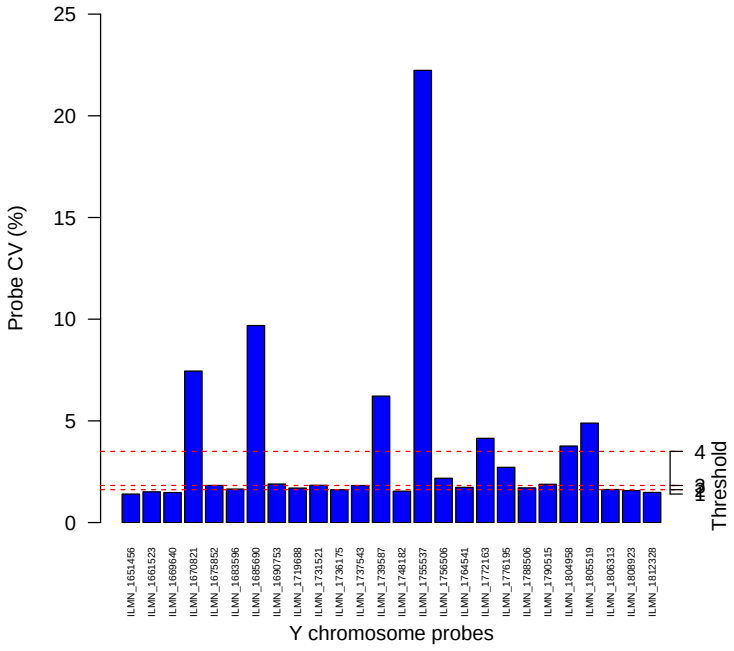

Figure 1: Expression variation (CV) of Y chromosome probes across all samples

At this data extraction step, you have the option of using you own list of probes corresponding to Y chromosome genes or using the probe lists included with the package. The included lists correspond to popular microarray platforms and contain identifiers for probes that map to non-recombining Y chromosome genes.

To load all the included probe lists:

```
> data(y.probes)
```

For further details on the included probe lists and corresponding platforms, refer to the MASSI manual.

After viewing the probe variation plot, a decision can be made regarding which probes to use in the clustering step. The MASSI package includes methods for selecting probe variation thresholds based on quantiles. The threshold can be determined by quantiles of probe variance (CV): 1=All probes, 2=Upper 75%, 3=Upper 50%, 4=Upper 25%. It is highly recommended that probe CV plot generated using the massi.y function be inspected to inform threshold choice (Figure 1). Default=3. If the total number of Y chromosome probes in the dataset is less than 10, consider using a threshold of 1 (all probes).

Once a probe threshold has been decided, run the massi.check function to extract the Y chromosome probe data and check that the optimum number of clusters in the data is 2. The optimum number of clusters in the dataset is estimated using the data from the most variable Y chromosome probes by calculating the average silhouette width of clusters as implimented in the "fpc" package (Henning, 2013).

```
> massi.check(exprs=MASSI.test.dataset, y.probes=MASSI.test.probes,
+             threshold=3)
```

```
[1] "Good, Optimum number of clusters is 2, proceed to massi.cluster"
```

As you can see from running massi.check on the test dataset, that the optimum number of clusters is 2. If the optimum number of clusters is >2, then the probe selection step should be reviewed. This may indicate that there are factors other than sex influencing the these probe values.

4 Predicting the sex of samples

To classify samples as either male or female, clustering is performed using the values from the subset of Y chromosome probes by implementing the partitioning around medoids algorithm to perform k-medoids clustering (Henning 2013), where samples are assigned to one of two clusters. The two clusters are then compared using the probe expression values across all samples in each cluster. Samples within the cluster featuring the highest Y chromosome probe values are classed as male and those amongst the cluster with the lowest Y probe values classed as female. Results such sample probe mean, standard deviation and z-scores are reported in a table together with the sex predicted for each sample.

To predict the sex of the samples:

```
> results <- massi.cluster(y.subset.values=y.subset.values)

> head(results)

   ID mean.probe.value sample.sd rowMeans.z.score. kmedoids.sex
1  S1         6.104126  1.553569        -0.5720338       female
2 S10         6.569614  1.456682         0.4765851         male
3 S11         5.974334  1.512225        -0.8334588       female
4 S12         6.605231  1.517065         0.8191216         male
5 S13         6.070618  1.472386        -0.4632365       female
6 S14         6.042864  1.474384        -0.5774843       female
```

As you can see, this function generates a table with the predicted sex of each sample, along with other metrics.

5 Visualising the results

The results of the MASSI analysis can easily be visualied using the massi.plot function and the data values derived from the previous steps. To run the massi.plot function:

```
> massi.plot(y.values=y.values,
+            y.subset.values=y.subset.values, massi.output=massi.output)

pdf
  2
```

This will produce a plot of Y chromosome probe expression variance represented as the coefficient of variation, with a line representing hte selected threshold similar to Figure 1. A bar plot of mean values from the subset of Y chromosome probes used in K-medoids clustering (Figure 2), with the bar colours representing clusters with female as red and male as green. This wil also generate a heatmap with dendrogram of Y chromosome probes as rows and individual samples in columns (Figure 3) and a principal component plot showing clusters (Figure 4). All of these figures are saved to a file named "massi.figures.pdf".

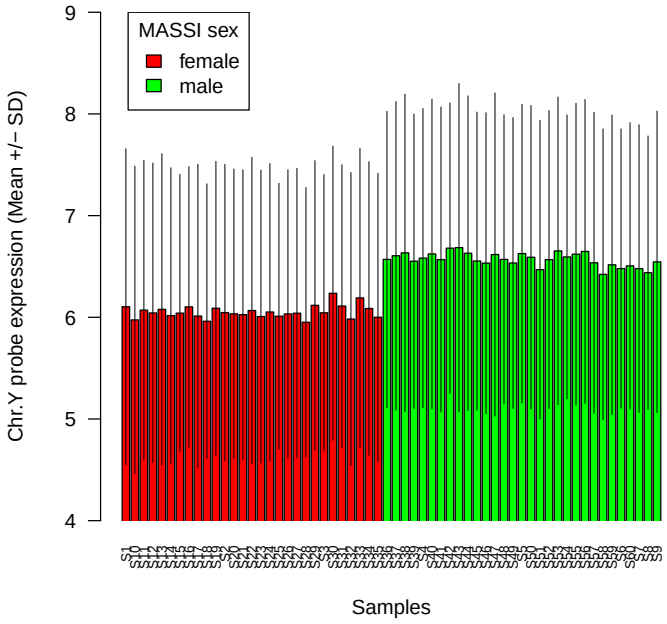

Figure 2: Mean values of the subset of Y chromosome probes used in K-medoids clustering. The bar colours represent clusters, which were assigned as female (red) and male (green)

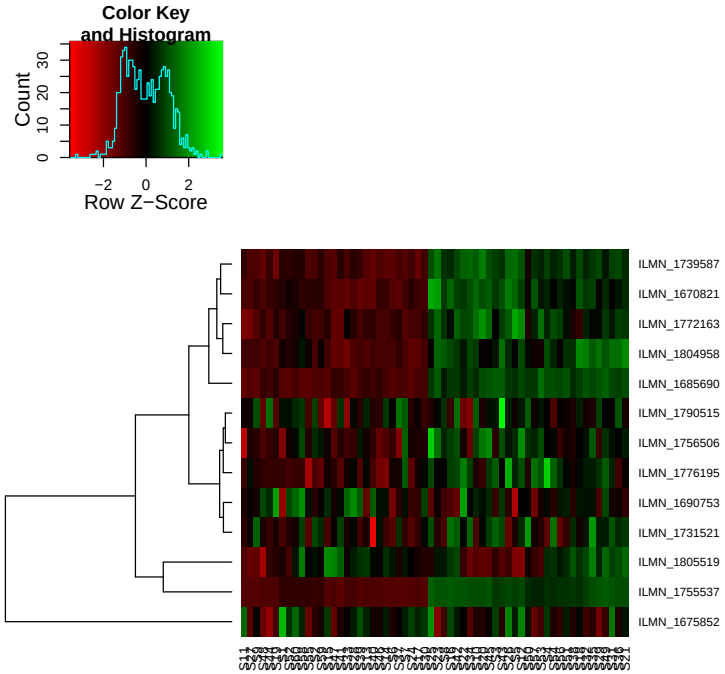

Figure 3: Heat map with dendrogram of Y chromosome probes as rows and individual samples in columns

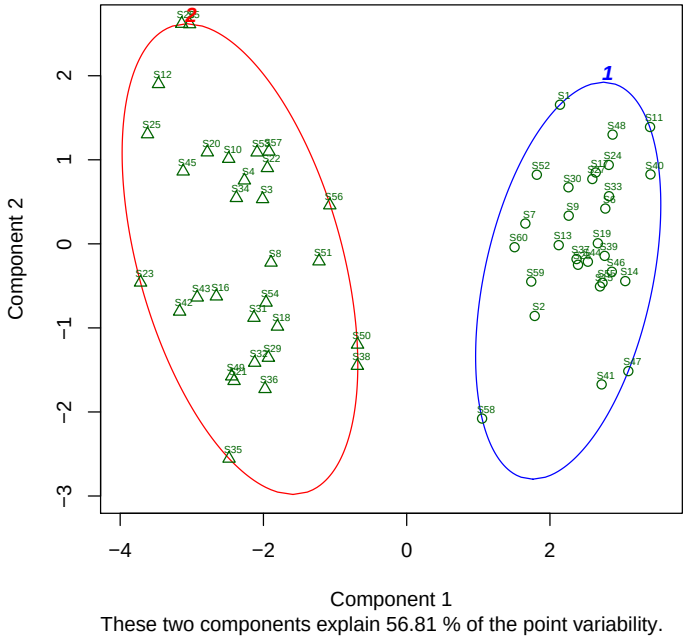

Figure 4: Principal component analysis plot of male and female clusters

6 Using the MASSI espresso function

All the key features in the MASSI package are incorporated into the massi.espresso function that uses default values. To use the massi.espresso function with the test data:

```
> data(MASSI.test.dataset)
> data(MASSI.test.probes)
> massi.espresso(exprs=MASSI.test.dataset, y.probes=MASSI.test.probes)
```

7 References

Henning, C. (2013) fpc: Flexible procedures for clustering. R package version 2.1-5. http://CRAN.R-project.org/package=fpc