Supplementary Information

massiR: a method for predicting the sex of samples in gene expression microarray datasets

Sam Buckberry[1,*], Stephen J Bent[1], Tina Bianco-Miotto[1,2] and Claire T Roberts[1]

[1] The Robinson Institute, School of Paediatrics & Reproductive Health, The University of Adelaide, Australia.

[2] School of Agriculture Food & Wine, The University of Adelaide, Australia

1 Y CHROMSOME PROBE INDENTIFIERS INCLUDED WITH THE MASSI PACKAGE

To identify probes that represent Y chromosome genes, we used the Ensembl mappings of probes for commercially available microarray platforms. We selected this option because Ensembl have independently mapped the probes from numerous platforms to a common reference genome, and the annotation information for many platforms is accessible through the Bioconductor package *biomaRt*. This method allowed us to select probes that map uniquely to Y chromosome genes. A detailed example on how to obtain probe information for commercial microarray platforms is included with the *massiR* vignette. For details on probe mapping methods, see the permalink: http://jan2013.archive.ensembl.org/info/docs/microarray_probe_set_mapping.html.

2 TESTING AND VALIDATION

We searched the NCBI GEO public repository for gene expression microarray datasets with associated sex information in the metadata for testing purposes (Supplementary Table 1). When raw data were available, we preprocessed and normalized the arrays before performing quality assessments using standard methods and Bioconductor packages in R. Any arrays that were deemed to be outliers were removed from the dataset, then the data were re-normalised before predicting the sex of samples using the *massiR* package.

To test the accuracy of this method, we selected ten datasets encompassing multiple microarray platforms and samples derived from various normal and pathological tissues (Supplementary Table 1). In 6/10 datasets, this method predicted the sex of the samples with 100% accuracy (Supplementary Table 1). However, this validation methodology is dependent on the accuracy of the associated metadata. Given that this prediction method only uses information from Y chromosome probes, we interrogated each dataset to examine probe-specific expression values for each sample to further understand why we encountered a few isolated cases of misclassification (see below). Therefore it is reasonable to suggest that some of these discrepancies may be due to unintended errors in the metadata and not due to misclassification.

2.1 Samples classified as male but listed as female in the metadata

There were four cases across two datasets where samples were predicted as male using this method but listed as female in the metadata (Supplementary Table 1). When interrogating the individual Y chromosome probe values, we observed that all of these samples show expression of Y chromosome genes well within the range of all the other male samples in the dataset (Supplementary Figures 1 & 2).

2.2 Samples classified as female but listed as male in the metadata

There were eight cases across three datasets where samples were classified as female using this method but indicated as male in the metadata (Supplementary Table 1). In all but one of these cases (Supplementary Figure 3) the intensity values for Y chromosome probes was well within the range of female samples, and showed no indication of any Y chromosome gene expression (Supplementary Figures 2-4). However, although this infers that several of these samples are female (as predicted), one cannot exclude the possibility that the cells assayed were not expressing Y chromosome genes at that time point.

2.3 Performance with skewed sex ratios

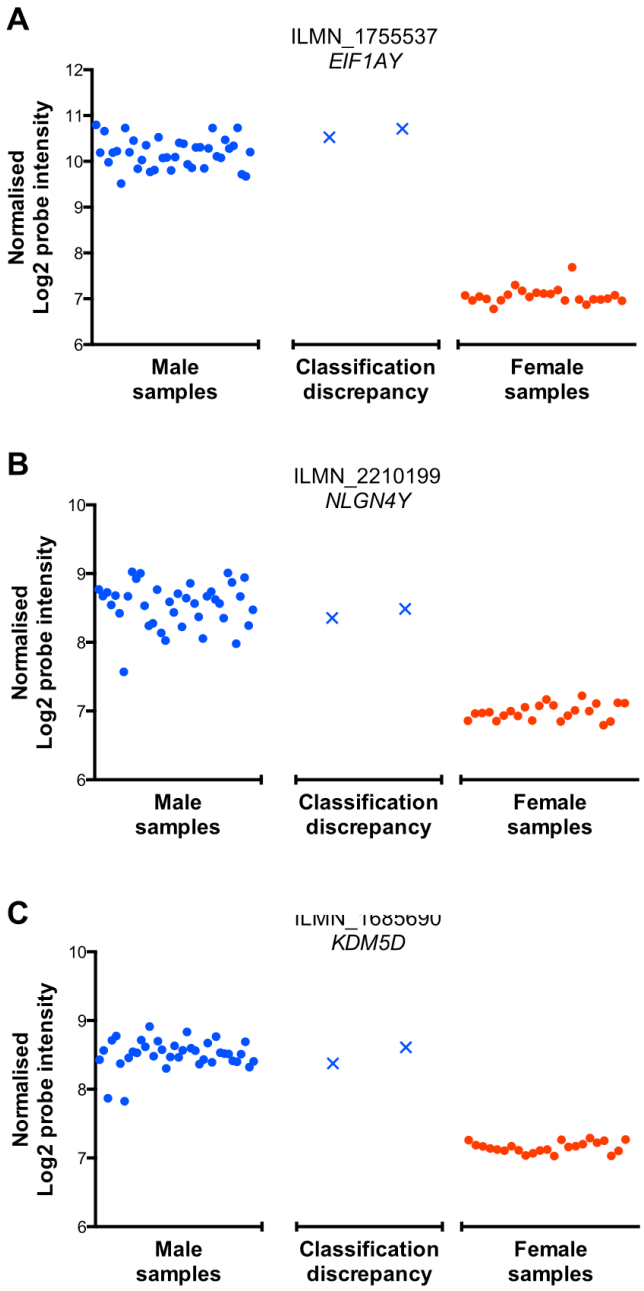
To test the performance of the *massiR* method with datasets with skewed sex ratios, we randomly selected male and female samples from large array datasets to generate random data subsets with a spectrum of sex ratios. This was performed with human brain (GSE29378), colorectal (GSE35896), kidney (GSE40435), placenta tissue (GSE25906) and peripheral blood mononuclear cells (GSE45330). For each dataset, we separated the male and female samples and then randomly selected samples from each group to create datasets of pre-determined sex ratios and sample sizes. For each dataset, we performed this randomized dataset construction process in triplicate. The summarized results for each tissue type are presented in Supplementary Figure 5.

2.4 Detecting datasets with skewed ratios.

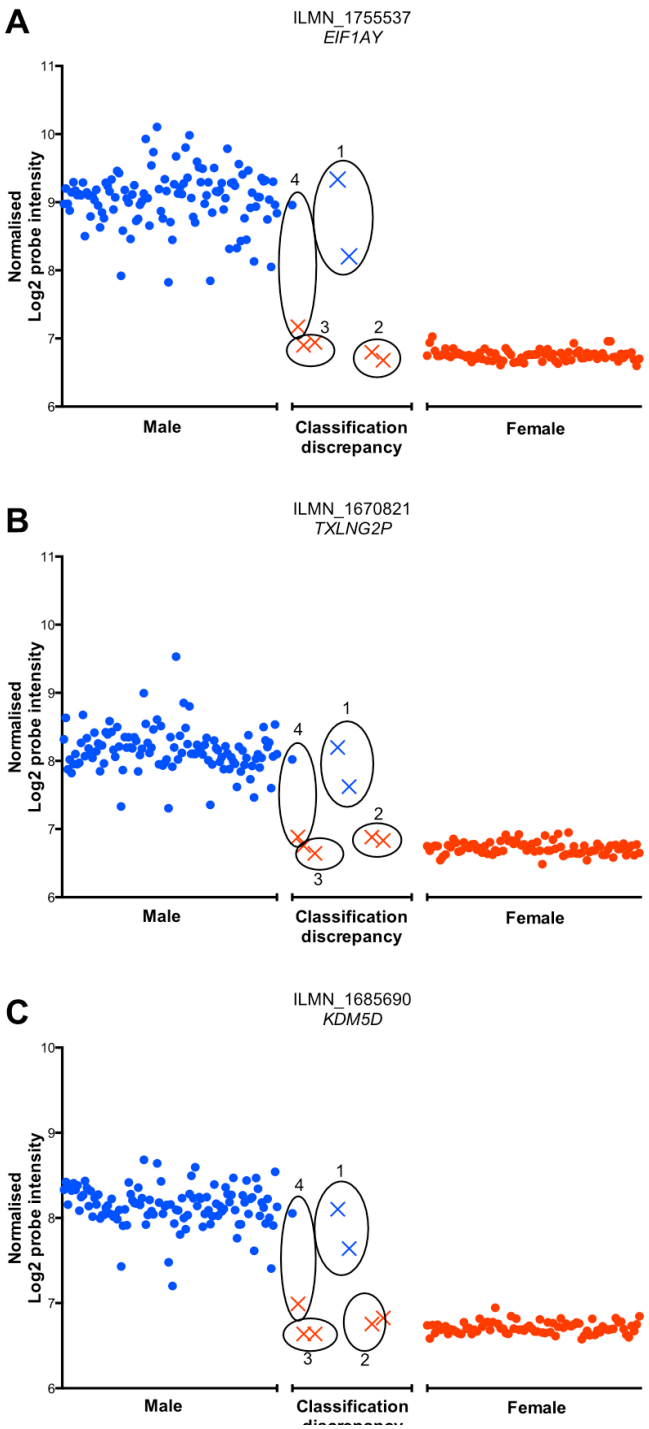
The *massiR* package includes a function that aids in detecting if a dataset has a skewed male/female ratio. This function calculates a standardized score for each sample and implements the dip test to test for unimodality. As a relatively sex-balanced dataset would typically show a bi-modal distribution of these standardized scores, the dip statistic is used to predict if a dataset shows a unimodal distribution that would be expected if a vast majority of samples were of one sex. We tested this function using the same randomly generated data as above to develop the guidelines for detecting dataset with skewed sex ratios which are outlined in the *massiR* package vignette (Supplementary Figure 6).

Supplementary Table 1. Validation results for predicting the sex of samples in microarray datasets using the MASSI package.

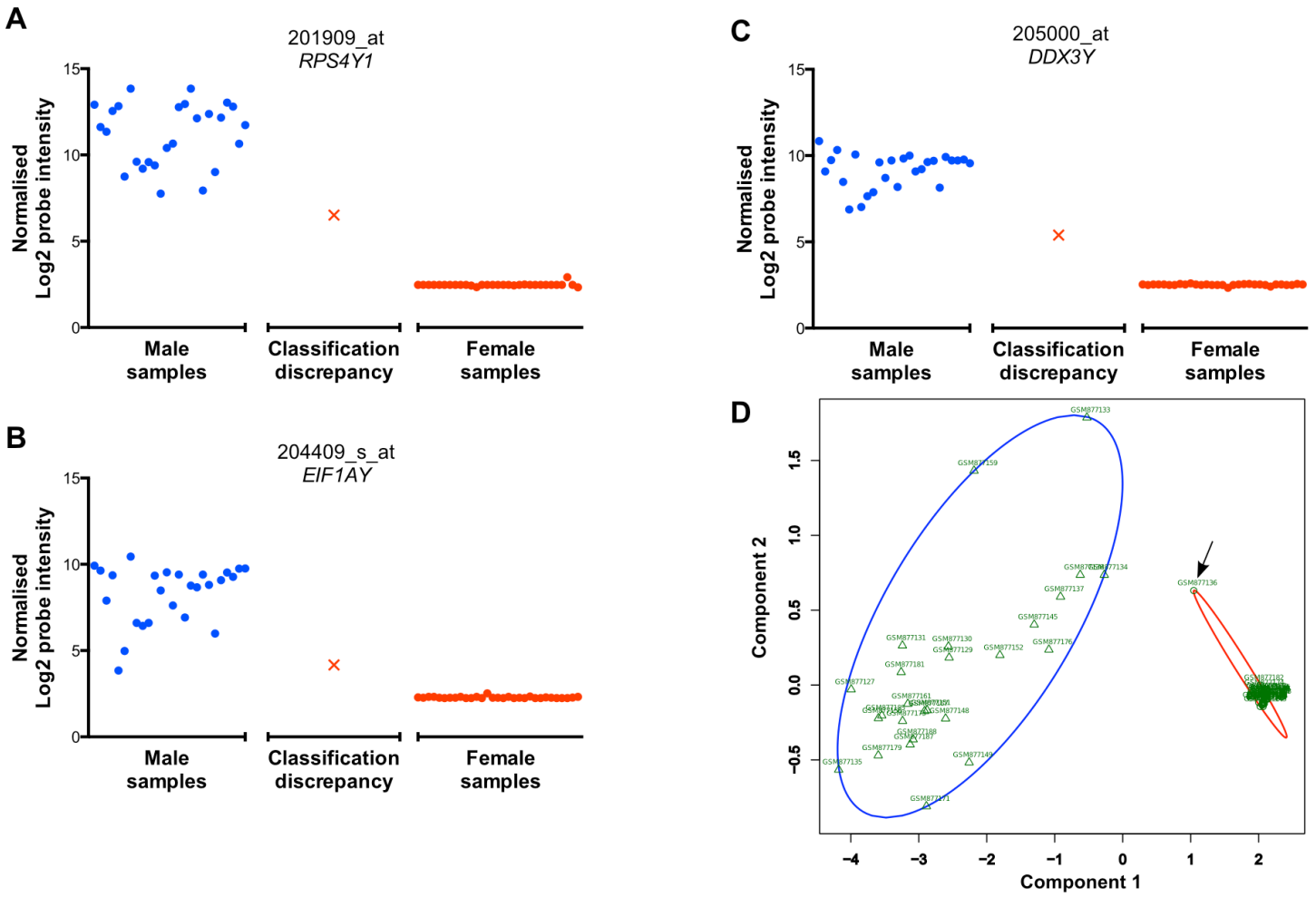
GEO accession	Species	Tissue	Platform	No. Samples	No. correctly predicted	Overall Prediction accuracy	Male samples			Female samples		
							No. samples	No. correctly predicted	Prediction accuracy	No. samples	No. Correctly predicted	Prediction accuracy
GSE45330	Human	Blood cells	Illumina Human HT-12 V4	77	76	**98.70%**	33	32	96.97%	44	44	100.00%
GSE29378	Human	Brain	Illumina Human HT-12 V3	63	61	**96.83%**	38	38	100.00%	25	23	92.00%
GSE35896	Human	Colorectal	Affymetrix HG-U133 Plus 2.0	58	57	**98.28%**	27	26	96.30%	31	31	100.00%
GSE25906	Human	Placenta	Illumina Human-6 V2	60	60	**100.00%**	31	31	100.00%	29	29	100.00%
GSE13546	Human	Lung cancer	Affymetrix HG-U133 Plus 2.0	15	15	**100.00%**	3	3	100.00%	12	12	100.00%
GSE20950	Human	Adipose	Affymetrix HG-U133 Plus 2.0	39	39	**100.00%**	12	12	100.00%	27	27	100.00%
GSE14335	Human	Fibroblast cells	Affymetrix HG-U133A 2.0	10	10	**100.00%**	3	3	100.00%	7	7	100.00%
GSE40435	Human	Kidney	Illumina Human HT-12 V4	202	195	**96.53%**	118	113	95.76%	84	82	97.62%
GSE29585	Mouse	Placenta	Affymetrix Mo. Exon 1.0 ST	16	16	**100.00%**	8	8	100.00%	8	8	100.00%
GSE35182	Mouse	Heart	Affymetrix Mo. Gene 1.0 ST	24	24	**100.00%**	12	12	100.00%	12	12	100.00%
Totals				**564**	**553**	**98.05%**	**285**	**278**	**97.54%**	**279**	**275**	**98.57%**



Supplementary Figure 1. Prediction and validation of the sex of samples in dataset GSE29378. Plots show microarray probe intensity values for Y chromosome genes *EIF1AY* (A), *NLGN4Y* (B) and *KDM5D* (C). This shows that two samples (blue crosses) listed as female in the metadata show Y chromosome gene expression values comparable to male samples (blue dots), which are distinct from the samples confirmed as female (red dots).

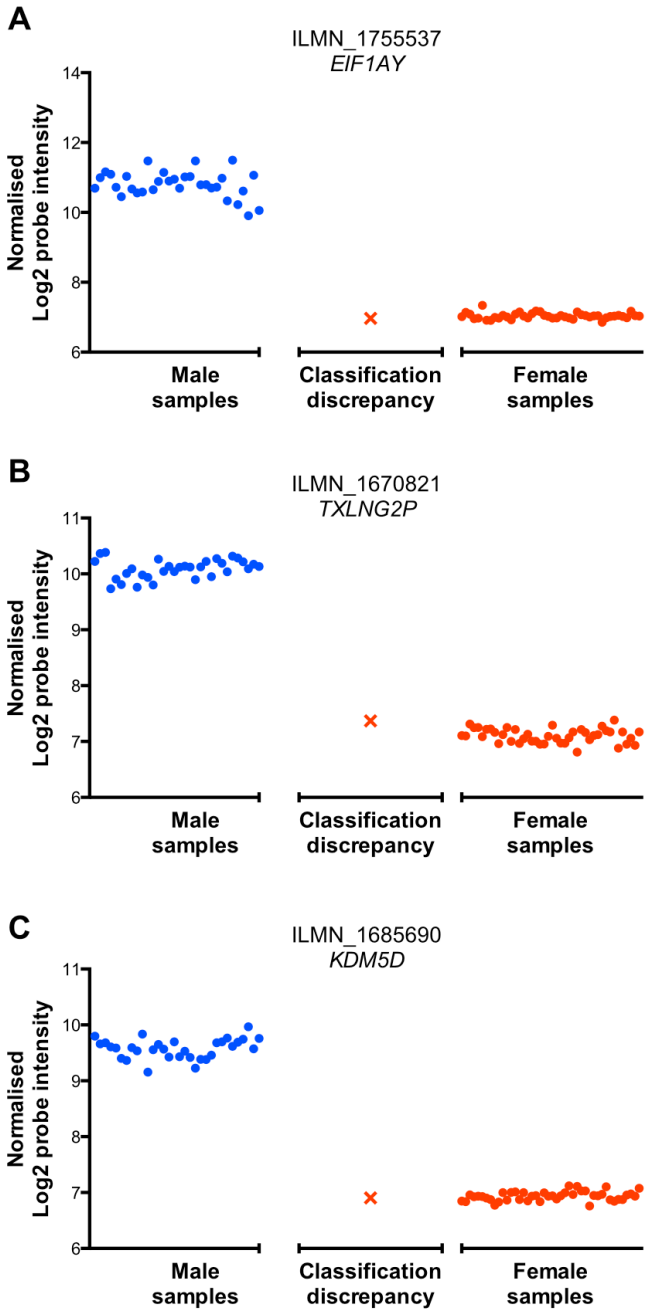


Supplementary Figure 2. Prediction and validation of the sex of samples in dataset GSE40435. Plots show microarray probe intensity values for Y chromosome genes *EIF1AY* (**A**), *TXLNGP2* (**B**), *KDM5D* (**C**). Samples are derived from paired tumor and adjacent non-tumor tissue. Dots within the male (blue) and female (red) groups were predicted to be the same sex as listed in the metadata. Samples with discrepant classification are represented by crosses, with the colour corresponding to the predicted sex. The pairs of samples within circle (1-4) were obtained from the same individual. These plots show the misclassification occurred for both samples in three pairs (1-3). In paired group 4, the misclassified sample was derived from the tumor tissue and the correctly classified sample was derived from adjacent normal tissue.

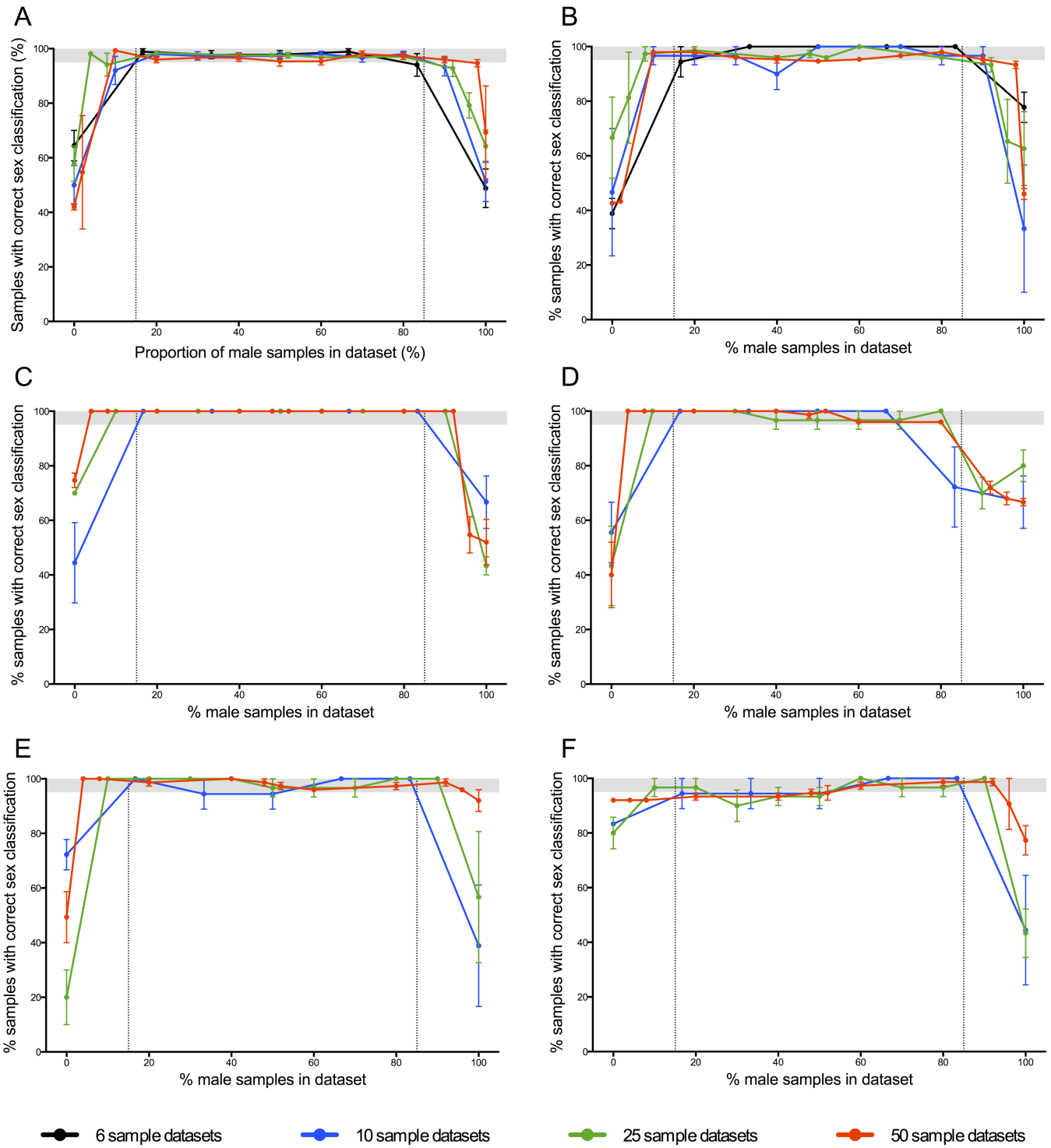


Supplementary

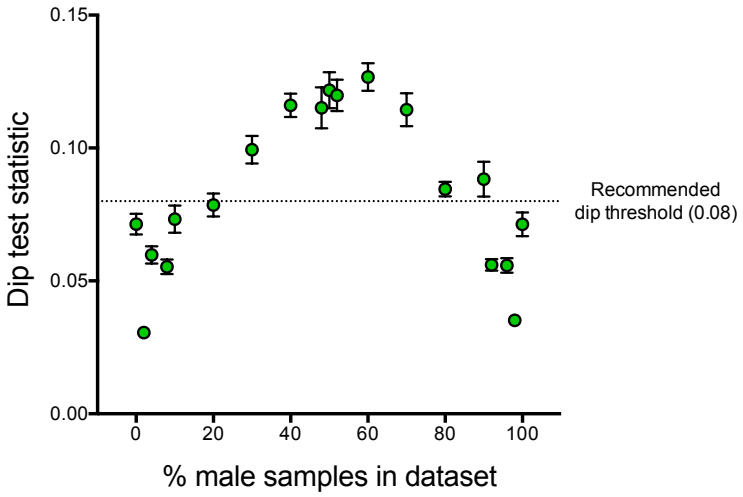
Figure 3. Prediction and validation of the sex of samples in dataset GSE35896. Plots show probe intensity values for Y chromosome genes *RPS4Y1* (**A**), *EIF1AY* (**B**), *DDX3Y* (**C**). One sample, indicated by the red cross was predicted to be female but listed as male in the metadata. This misclassified sample showed probe intensity values for all three genes greater than all other female samples (red dots), but less than that of males (blue dots), which suggests a genuine misclassification. When inspecting the PCA plot of these samples (**D**), this misclassified sample is plotted distinctly apart from the other female samples, although placed within female cluster.



Supplementary Figure 4. Prediction and validation of the sex of samples in dataset GSE45330. Plots show probe intensity values for Y chromosome genes *EIF1AY* (**A**), *TXLNGP2* (**B**), *KDM5D* (**C**). One sample, indicated by the red cross was predicted to be female but listed as male in the metadata. The Y chromosome probe intensity values for this sample are in the range of all other female samples (red dots) in this dataset and distinct from all the male samples (blue dots).



Supplementary Figure 5. Sex prediction accuracy of the *massiR* package using five human gene expression datasets and a range of male/female ratios. **A** Results summary of the five datasets, **B** Kidney tissue (GSE40435), **C** placenta tissue (GSE25906), **D** colorectal tissue (GSE35896), **E** Blood mononucleocytes (GSE45330), **F** brain tissue (GSE29378). Points represent mean, vertical bars represent the standard error of the mean. The grey band at the top of the plot shows the 95-100% range. The correct sex prediction rate is 97.2% (±1.2 SEM) for datasets with >15% and <85% males which is the area between the vertical dotted lines.



Supplementary Figure 6. The dip test statistic as a method for identifying datasets with a skewed sex ratio. This plot shows the relationship between the dip test statistic as returned by the massi.dip function and the proportion of males in the dataset. This plot summarizes randomly selected sample and data subsets adapted from empirical kidney tissue (GSE40435), placenta tissue (GSE25906), colorectal tissue (GSE35896), Blood mononucleocytes (GSE45330), brain tissue (GSE29378) datasets. Points represent mean, vertical bars represent the standard error of the mean. Datasets with a dip test statistic greater than the threshold (0.08) are unlikely to feature skewed sex ratios that will affect the performance of predicting sample sex using the *massiR* package.