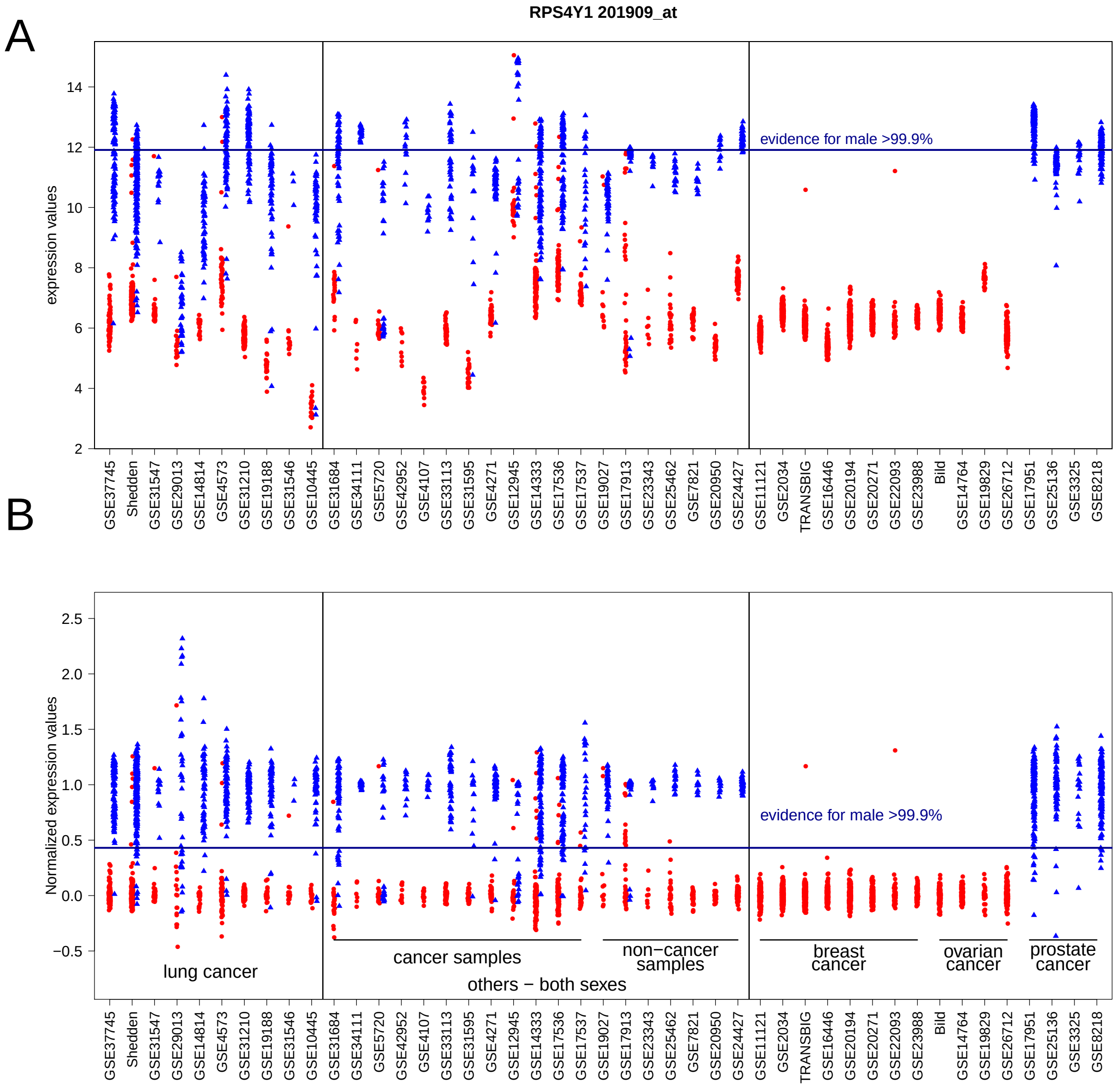
Supplemental figures of

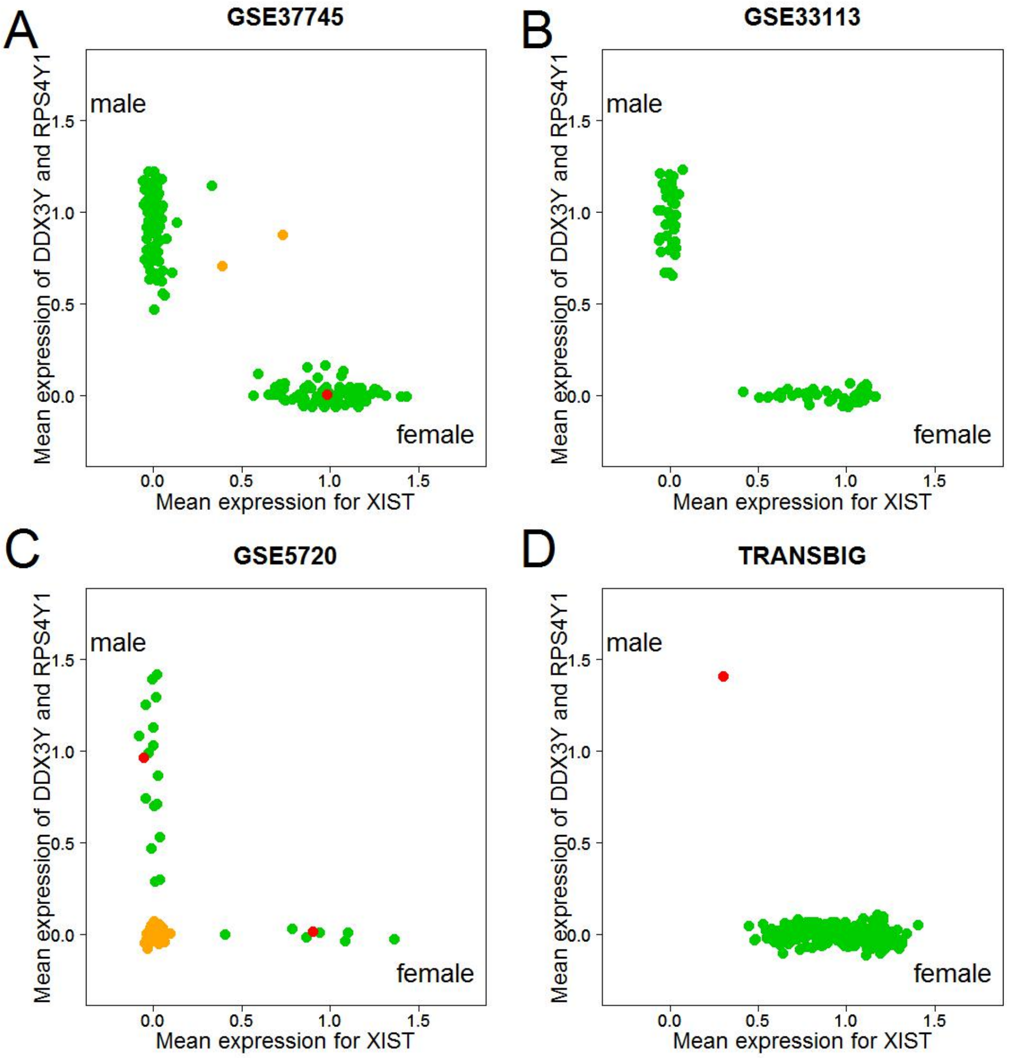
Miriam Lohr, Birte Hellwig, Karolina Edlund, Johanna Mattsson, Johan Botling, Marcus Schmidt, Jan G. Hengstler, Patrick Micke, Jörg Rahnenführer:

Identification of sample annotation errors in gene expression datasets

Supp. Fig. 1: Improvement of comparability of cohorts by normalization. (A) Raw expression values of female (red) and male (blue) labelled samples for probe set 201909_at (*RPS4Y1*) across all datasets. (B) The same cohorts after normalization. Specifically, two outliers in datasets TRANSBIG and GSE22093 indicate two breast cancer patients with high *RPS4Y1* expression, a feature clearly inconsistent with female sex.



Suppl. Fig. 2: Visualization of the male-female classifier with mean expression values of the two probe sets for *XIST* on the x-axis and of *DDX3Y* and *RPS4Y1* on the y-axis. The points represent individual patients. The point clouds on the left and on the bottom are characteristic for males and females, respectively. Colors indicate classification accuracy of samples. Green: "correctly classified", red: "misclassified", and orange: "unconfident". A. Results for the Uppsala cohort (GSE37745): One female patient is clearly mislabeled as male, two samples are labeled "unconfident". B. Results for GSE33113 with clear discrimination between males and females and no sex misannotations. C. Results for GSE5720 with two misclassified samples and a large number of samples classified as "unconfident". D. Results for a breast cancer dataset (TRANSBIG) with one male patient assigned to the category "misclassified".


A
GSE37745
B
GSE33113
C
GSE5720
D
TRANSBIG
male
female
Mean expression of DDX3Y and RPS4Y1
Mean expression for XIST