Description of the algorithms used for building and applying the sex classifier described in

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

Overview:

1. Algorithm for selection of suitable variables (genes, probe sets) for sex classification
2. Algorithm for normalization of datasets
3. Algorithm for the sex classifier

Algorithm 1: Selection of variables (genes, probe sets) for sex classification

Input:
- Raw expression values for l cohorts (training datasets) for a list of candidate genes
- Sex annotation for every sample (male/female)
- Cutoff q for probability for sex evidence
- Cutoff c for median classification accuracy per gene

Output: List of genes for classifier with classification accuracy above c

Algorithm: All steps are carried out for each gene separately. First carry out steps 1-4 for every dataset ($i = 1, \ldots, l$) separately, then steps 5-6.

Step 1 Estimate location of the values for males and for females, respectively, with the robust estimate *median*.

Step 2 Call the smaller median m_0 and the larger m_1, and estimate the scale with the robust Rousseeuw-Croux estimator $Q_{n,0}$ of the lower expression values.

Step 3 Assign a sex-specific Gaussian distribution f_0 to the low expression values, with parameters $\mu_0 = m_0$ and $\sigma_0^2 = (2.22\, Q_{n,0})^2$.

Step 4 Calculate the q quantile of the distribution f_0 and classify a sample to the group with larger values if its expression value is above this cut point. Count the fraction of correctly classified samples, where the label is male (female) and the corresponding expression value belongs to the expression values of the male (female) group.

Step 5 Compute the median of correctly classified samples across all datasets and call it *median classification accuracy*.

Step 6 If the *median classification accuracy* is above the cutoff c, include the corresponding gene in the sex classifier.

Algorithm 2: Normalization of values across cohorts

Input:

- Raw expression values for a set of cohorts and for a specific gene
- Assignment of cohorts to training datasets and test datasets
- Sex information (if cohort belongs to training datasets)

Output:

Normalized expression values with group medians 0 and 1 for sex-specific subgroups, for every cohort.

Algorithm:

If the samples belong to both sexes, proceed as follows:

Step 1 Calculate sex-specific location measures.

- For a training dataset compute sex-specific medians and call the smaller value m_0 and the larger value m_1.
- For a test dataset (or if no sex assignment is available) cluster all values with the k-means algorithm into 2 groups, choose as starting values the 25% and 75% quantile of all values. Call the smaller value of the two resulting cluster centers m_0 and the larger value m_1.

Step 2 Calculate normalized expression values $\tilde{x}$ from original values x by

$$\tilde{x} = \frac{x - m_0}{m_1 - m_0}.$$

If all samples belong to one sex only (e.g. only females in a breast cancer dataset or only males in a prostate cancer dataset) proceed as follows:

The algorithm is described for a dataset d of females, for males proceed analogously.

Step 1* Calculate the normalized expression values for all training datasets as described in **Step 1**. Then estimate a typical variation for females as the value of $Q_{n,\text{female}}$, applied to the set of all normalized values of females across all training datasets.

Step 2* Estimate location and scale of the values in d with the robust estimates $m_d = median_d$ and $Q_{n,d}$.

Step 3* Calculate normalized expression values $\tilde{x}$ from original values x in d by

$$\tilde{x} = \frac{x - m_d}{Q_{n,d}} Q_{n,\text{female}}.$$

Step 4* Add 1 to the normalized values if females correspond to the group with larger values, for the considered gene.

Algorithm 3: Sex classification

Input:
- Normalized expression values $\tilde{x}$ for one cohort d for a list of p genes, as a result of Algorithm 2
- Normalized expression values with group medians 0 and 1 for sex-specific subgroups for training datasets
- Chromosome allocation of the p genes (located on X or on Y chromosome)
- q Quantile q_d of the estimated distribution

Output: Classification of all samples of the cohort d as *correctly classified*, *misclassified*, or *unconfident*.

Algorithm: First carry out steps 1-2 for each of the p genes separately, then proceed with step 3.

Steps 1-2 are described for a gene located on the X chromosome, for genes on the Y chromosome proceed in the same way, but with roles of females and males interchanged.

Step 1 Assign a sex-specific Gaussian distribution f_0 to the normalized expression values for males (as in Step 3 of Algorithm 1), with parameters $\mu_0 = 0$ and $\sigma_0^2 = (2.22\, Q_{n,0})^2$, the latter based on all values that correspond to samples clustered to the male group, from all training datasets.

Step 2 Compare each normalized expression value $\tilde{x}$ with the q quantile of the estimated distribution q_d, and set the *female evidence score* to 1, if the normalized expression value $\tilde{x}$ is above the quantile q_d, and to 0 otherwise.

Step 3 Classify a sample as

- *correctly classified*, if at least one *female evidence score* is 1, all *male evidence scores* are 0, and the sample was labeled as female; or analogously for interchanged sex roles.
- *misclassified*, if at least one *male evidence score* is 1, all *female evidence scores* are 0, and the sample was labeled as female; or analogously for interchanged sex roles.
- *unconfident*, in all other cases.