EDA of RNA-Seq pseudocounts from Jaffe *et al.*

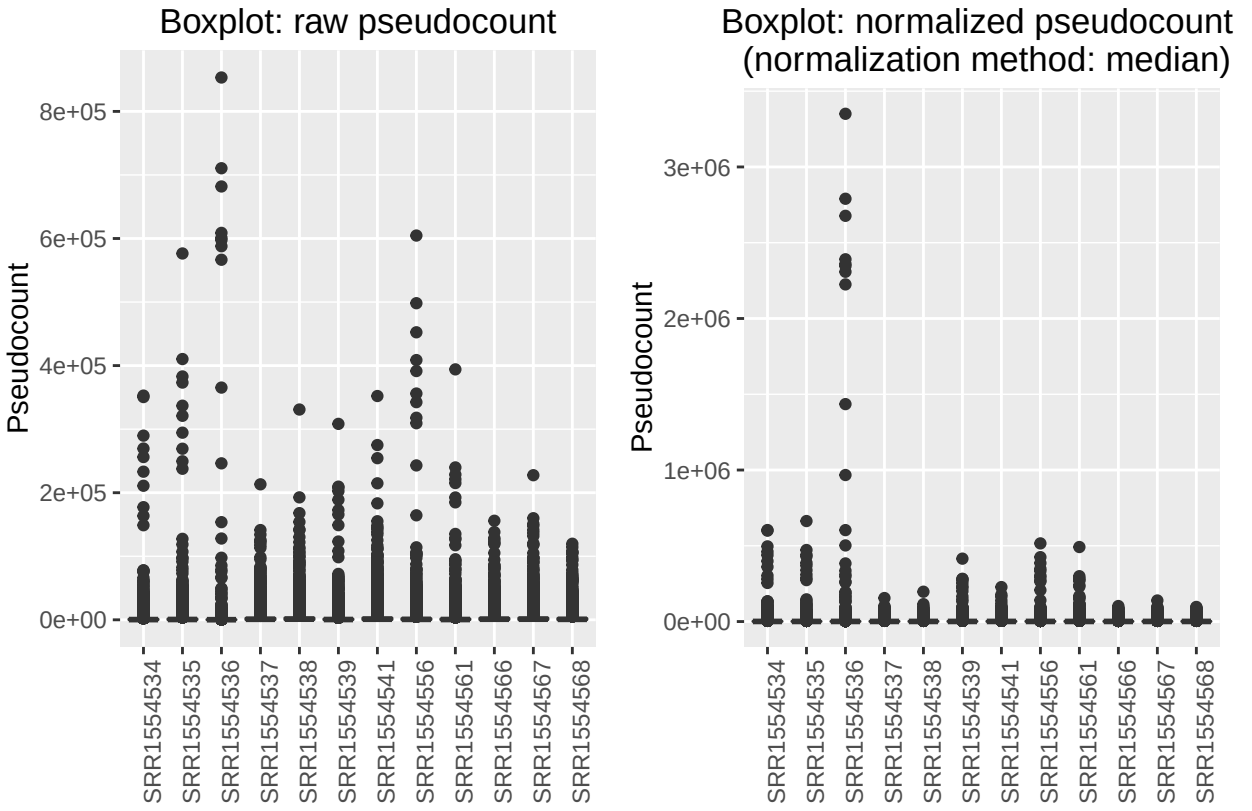
Nima Hejazi

2016 Oct 30 (Sun), 21:08:43

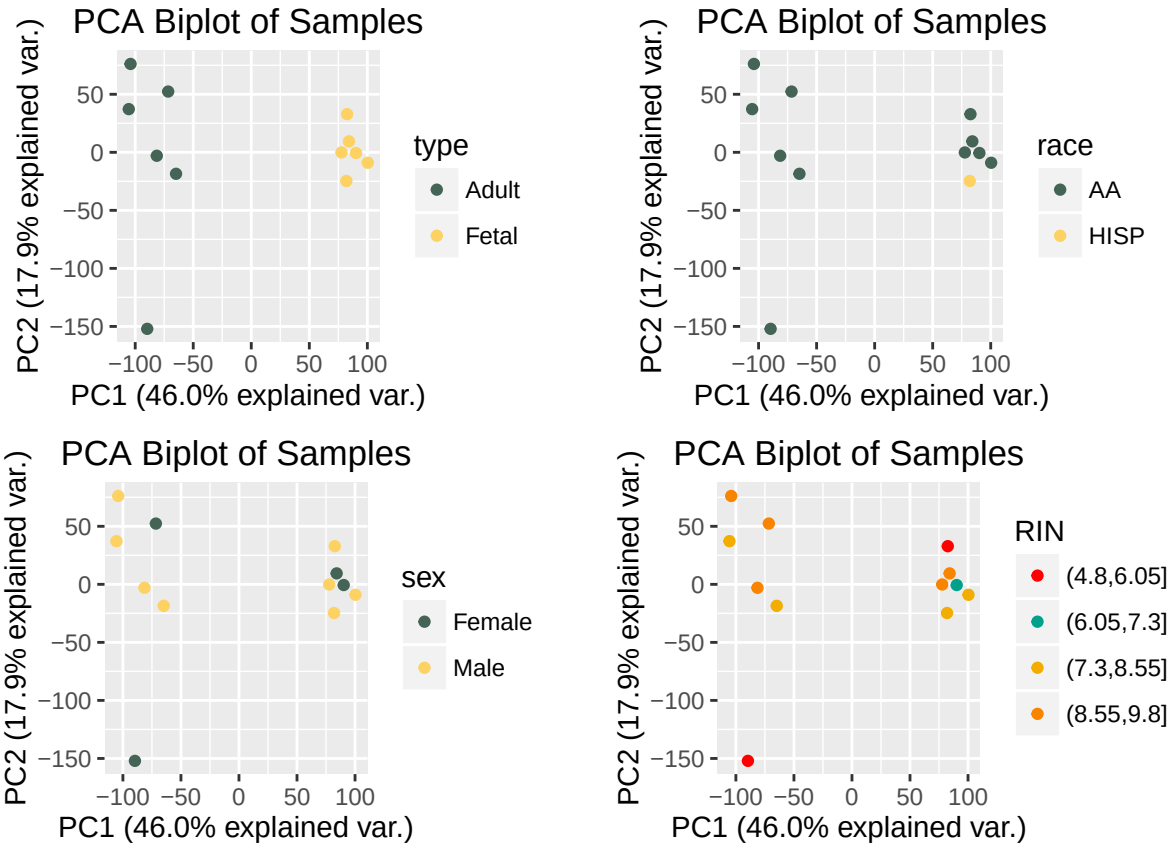
```
## Your platform/environment has not detected OpenMP support. fwrite() will still work, but slower in s
## Your platform/environment has not detected OpenMP support. fwrite() will still work, but slower in s
```

We provide several plots produced as part of the exploratory data analysis (EDA) procedures performed in examining this data set, and comment on the information provided by the EDA procedure.

Note that the plots are generated from a matrix of "pseudocounts," created from the use of the pseudoalignment algorithm proposed in Bray *et al.* (2016). Since the alignment procedure differs slightly from that used in standard alignment software (*e.g.*, bowtie), the counts are not integers. For any concerns about the validity of this alignment procedure, please consult the Bray *et al.* paper or the website of the Kallisto pseudoaligner.



The boxplot of the raw (pseudo)counts (graph on the left) indicates considerable variability across the twelve samples, with **SRR1554536** containing a number of (pseudo)counts that could be considered outliers. Among the other eleven samples, there is a fair amount of variability, though this degree of variation is generally expected and acceptable. Based on the empirical distributions of the raw (pseudo)counts, some form of normalization is warranted. To remove a significant degree of unwanted variation, we perform **median normalization**, a procedure which forces the medians of the distributions across samples to be the same. Careful examination of the normalized (pseudo)counts (graph on the right) suggests that the normalization procedure appears to force the empirical distributions to be similar, indicating that the procedure removes a deal of technical variability, while preserving variation that might be considered to be due to biological variability between samples. In spite of the normalization procedure, the empirical distribution of sample **SRR1554536** still appears to be problematic, suggesting that this sample may suffer from issues of quality and should likely be downweighted in later steps involving modeling.



The four principal component biplots presented above show a clear separation between the samples after projection into the space spanned by the first two principal components. A cursory examination of the biplot in the top left corner of the plot matrix indicates a clear separation between the samples based on the phenotype of interest (**fetal** vs. **adult**); furthermore, it is clear from the biplot that the samples are clearly separated in even the subspace of the first principal component (x-axis). The remaning principal component biplots indicate that the samples are fairly balanced across phenotype measures that are not directly of interest but may contribute to downstream analytic results. In particular, the biplot in the top right corner of the plot matrix indicates that nearly all of the samples originated from individuals of the same race, while the biplot in the bottom left corner of the plot matrix shows that each grouping of samples (across the main phenotype of interest) is balanced across sexes. The biplot in the bottom right corner of the plot matrix indicates that the considerable variation in quality of the samples (as measured by **RIN**) is fairly balanced across the two groups. On a final note, it is worth bringing attention to the fact that a single sample is separated from the rest after projection into the subspace of the first two principal components, and that it is this sample that corresponds to **SRR1554536**, which was noted to contain a number of outlying counts in the boxplots displayed in the previous section.