Project 1

Paulina Kucharewicz

January 9, 2023

1 Data exploration

a) There are 72208 observations and 5000 variables in the train data set, and 18052 observations and 5000 variables in the test data set.

b) As shown on Figures 1 and 2, almost all read counts are very low. Looking at histograms depicting clipped data (only read counts lower than 10) it is clear, that there is an abundance of zeroes.

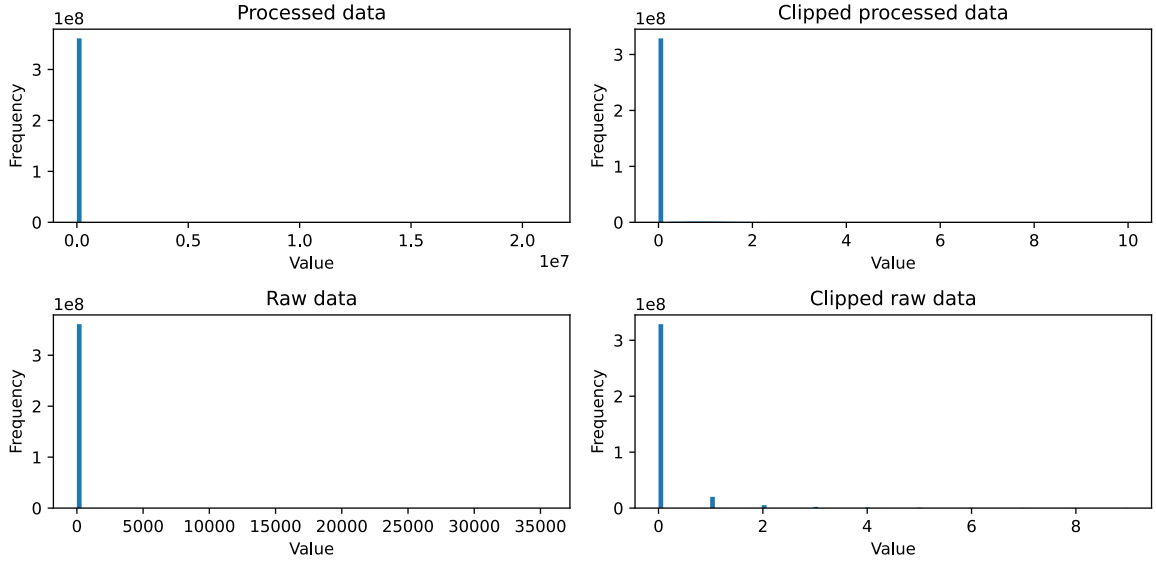

Figure 1: Histograms of train data (including zeroes). Clipped data consists of read counts lower than 10.

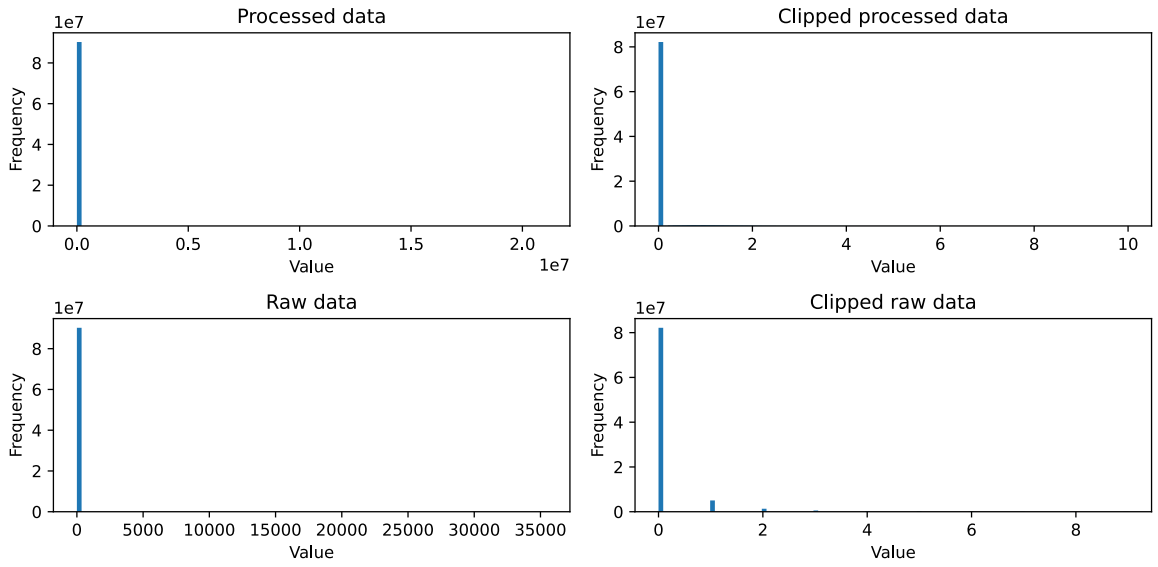

Figure 2: Histograms of test data (with zeros). Clipped data consists of read counts lower than 10.

c) Data was not normalised to 10k reads and was not log1p transformed. Data also was not scaled to unit variance.

d) There were histograms plotted after filtering out zeros, which are shown on Figures 3 and 4. After additionally clipping data (read counts lower than 10), more details of data distribution are visible.

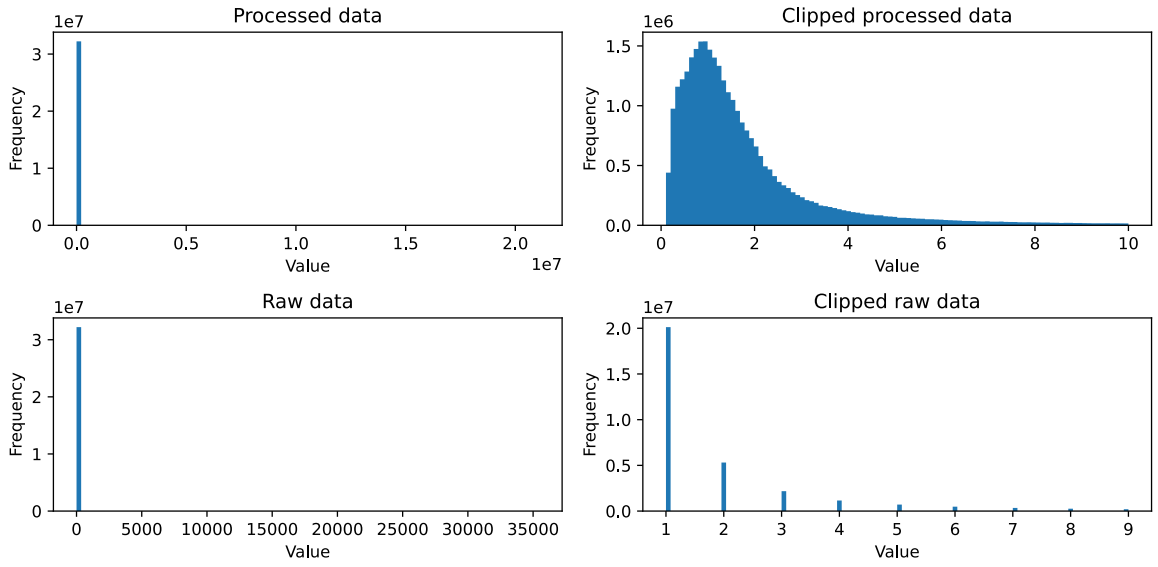

Figure 3: Histograms of train data with zeros filtered out. Clipped data consists of read counts lower than 10.

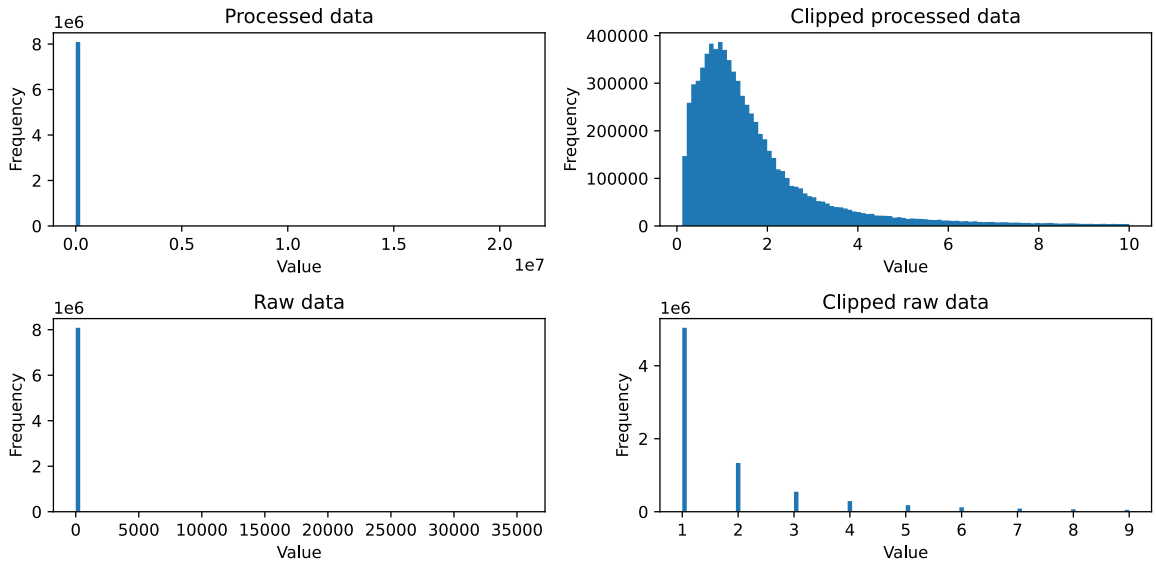

Figure 4: Histograms of test data with zeros filtered out. Clipped data consists of read counts lower than 10.

e) This data set has count data. Such data can be described with Poisson or negative binomial distributions. However, in Poisson distribution mean and variance are equal and for this data set these values are very different. Our data have negative binomial distribution, possibly zero-inflated. Abundance of zero can be explained by no transcription of genes or absence of mRNA (caused by faster mRNA degradation than transcription). Gene expression is different in different type of cells, some genes are very rarely expressed in general and some are not expressed in certain cell types. Many genes are also not expressed constantly, but only under certain conditions (gene regulation). All of the above can lead to elevated levels of read counts equal to zero.

f) Object `adata.obs` is an annotation of observations. Each row corresponds to one of the cells from experiment and contains information such as type of a cell, its donnor, batch etc.. From this data frame it can be concluded that there were:

- 9 patients (donors),
- 4 labs (sites),
- 45 cell types.