Curating RNAseq data for mouse and *Eimeria falciformis*

Emanuel Heitlinger

February 12, 2016

1 Objective

In the differential expression analysis of RNAseq data we assume a negative binomial distribution of transcript abundance. We analyse the distributions in mouse and parasite data respectively. A cutoff is applied to reads mapping per gene in raw data. For instance, a cutoff of ten requires that at least ten reads from *any sample* map to a given gene for the gene to count as expressed. We evaluate the effects of different cutoffs on the distribution.

2 Results for mouse

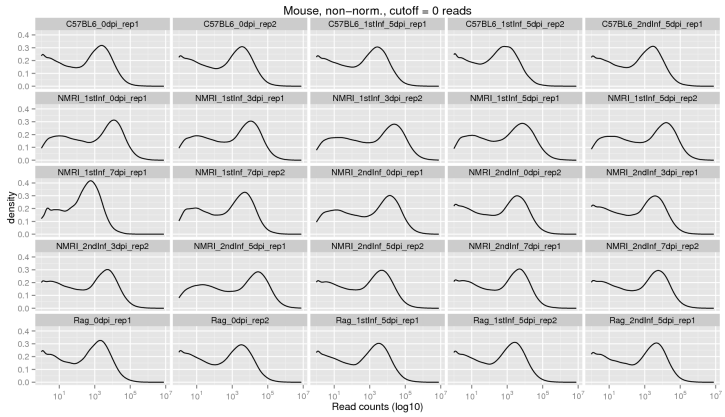

Figure 1: Density distribution of raw transcript counts (log10) for all mouse samples in analysis. All samples share the same bimodal distribution trend by visual inspection. The first density peak occur at $1 - 10^2$ transcripts, *i.e.* transcripts which are detected between 1 and 100 times in that sample. The second density peak occurs in the range of $10^3 - 10^4$.

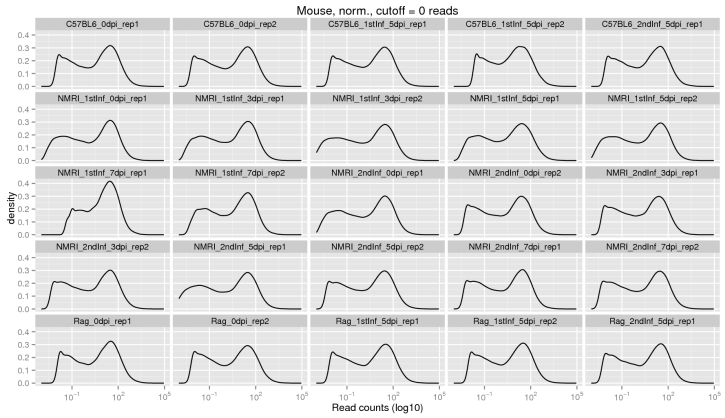

Figure 2: Density distribution of normalised transcript counts (log10) for all mouse samples. All samples share the same bimodal distribution trend by visual inspection. The first density peak occurs at < 1 transcripts, and in most samples it is more distinct than in the non-normalised data in Figure 1. For this normalised data, the second density peak occurs at roughly the same count value in all samples, close to 10^2 transcripts.

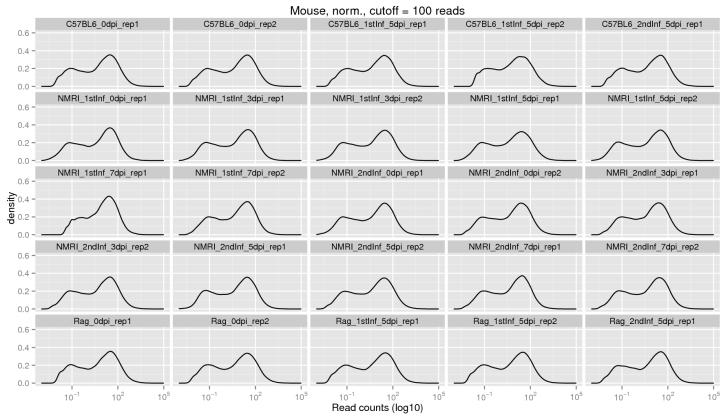

Figure 3: Density distribution of normalised transcript counts (log10) for mouse samples with a cutoff of 100 (see Methods for details). All samples still have a bimodal distribution, however less pronounced compared to figures 1 and 2. The first density peak occurs at < 1 transcripts as in Figure 2. The second density peak occurs at close to 10^2 transcripts and did hence not change from the normalised data without cutoff (Figure 2).

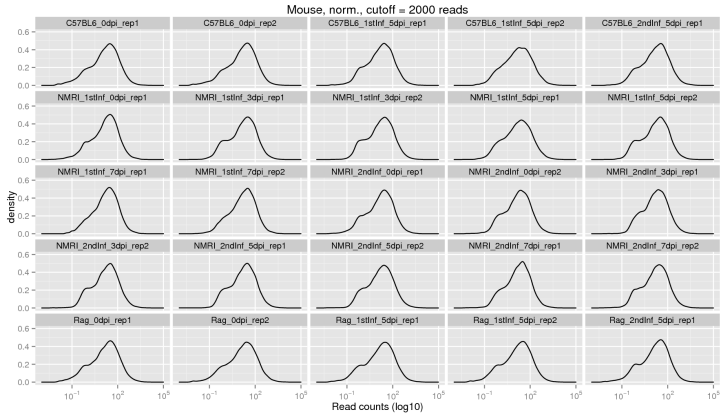

Figure 4: Density distribution of normalised transcript counts (log10) for mouse samples with a cutoff of 2000 (see Methods for details). The bimodality seen at lower cutoff values is strongly weakened or not detectable visually in some samples. It is however visually still seen in a majority of the samples. The first tendency to a density peak has shifted somewhat towards higher transcript counts (right) but is still < 1 transcripts as in Figure 3. The second density peak occurs at close to 10^2 transcripts and did hence not change from the normalised data with a lower cutoff (Figure 3).

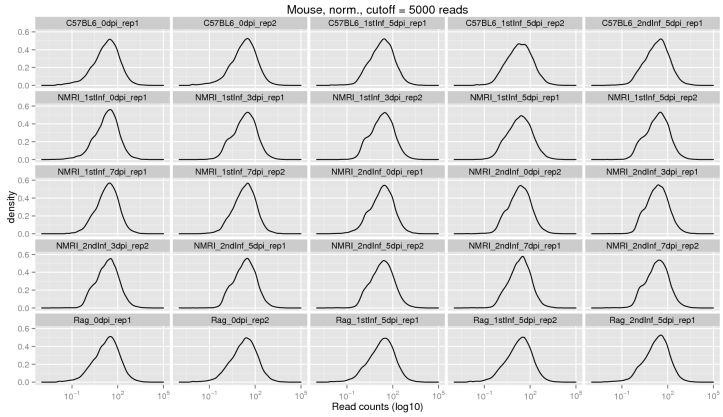

Figure 5: Density distribution of normalised transcript counts (log10) for mouse samples with a cutoff of 5000 (see Methods for details). The bimodality is not visually detectable in most samples, however it is still seen in *e.g.* samples NMRI_1stInf_3dpi_rep1 and NMRI_2ndInf_7dpi_rep2. The major density peak remains at a value close to 10^2 as in all the analysed normalised datasets (*e.g.* Figure 2 or 4).

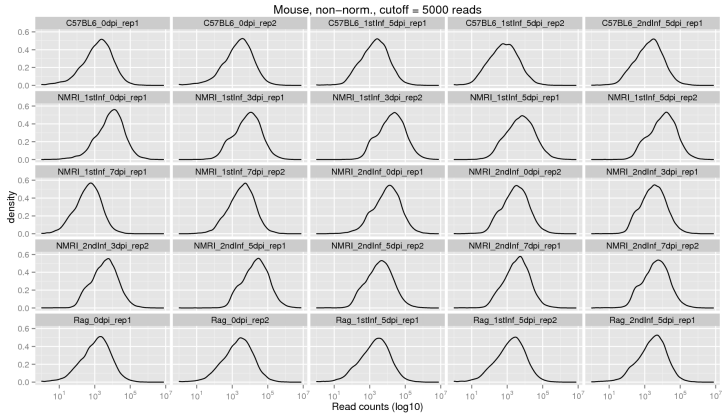

Figure 6: Density distribution of raw transcript counts (log10) for mouse samples with a cutoff of 5000 (see Methods for details). The bimodality is not visually detectable in most samples, however it is still seen in *e.g.* samples NMRI_1stInf_3dpi_rep2 and NMRI_2ndInf_7dpi_rep2. The major density peak is seen at a value over 10^2, shifting it almost one order of magnitude compared to the normalised data with the same cutoff (Figure 5).

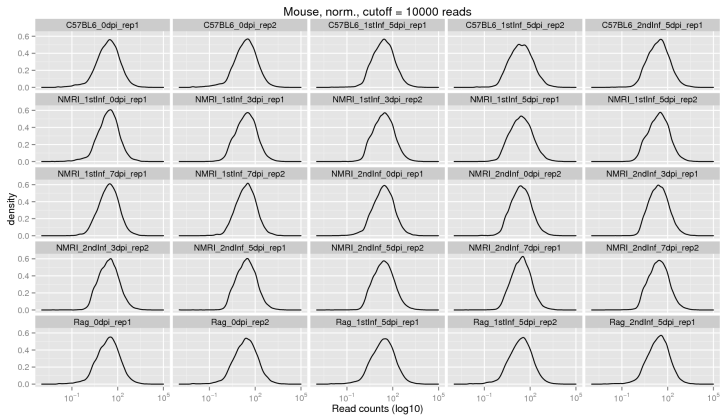

Figure 7: Density distribution of normalised transcript counts (log10) for mouse samples with a cutoff of 10000 (see Methods for details). The bimodality is not visually detectable in any sample apart from a minor indication in sample NMRI_1stInf_5dpi_rep2. The major density peak remains at a value close to 10^2 as in all the analysed normalised datasets (*e.g.* Figure 2 or 4).

3 Results for parasite

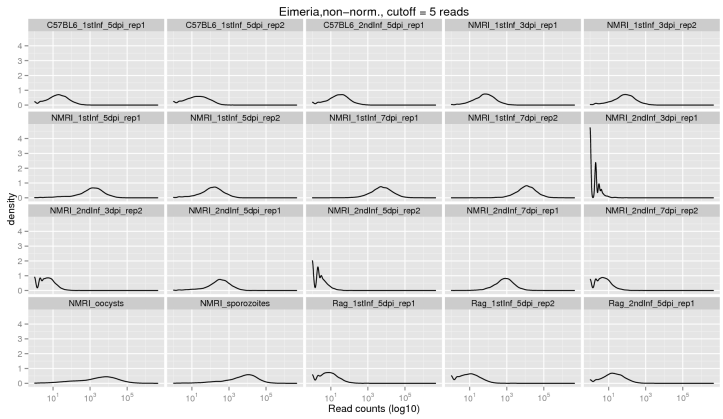

Figure 8: Density distribution of raw transcript counts (log10) for parasite samples with a cutoff of 5 (see Methods for details). Upon visual inspection distributions in most samples appear smooth and not contradictory to a negative binomial distribution. Density peak positions with regards to transcript counts (x-axis) are however at different orders of magnitudes between samples. Samples NMRI_2ndInf_3dpi_rep1, NMRI_2ndInf_3dpi_rep2, NMRI_2ndInf_5dpi_rep2 and NMRI_2ndInf_7dpi_rep2, however, do not.

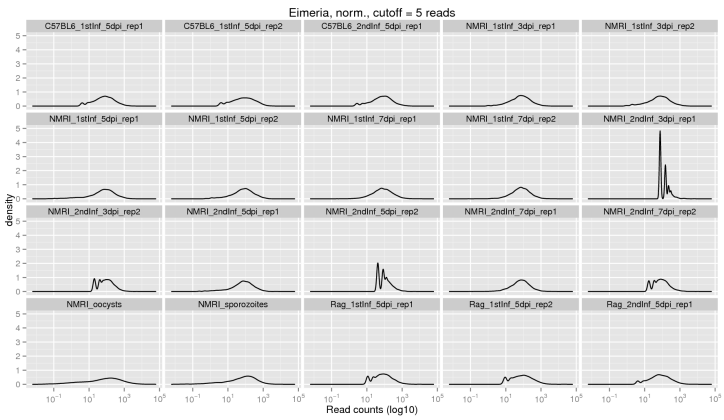

Figure 9: Density distribution of normalised transcript counts (log10) for parasite samples with a cutoff of 5 (see Methods for details). Upon visual inspection distributions in most samples appear smooth also in the normalised data. Density peaks are also in the same order of magnitude with regards to transcript counts (x-axis). Non-negative binomial samples in Figure 8 (NMRI_2ndInf_3dpi_rep1, NMRI_2ndInf_3dpi_rep2, NMRI_2ndInf_5dpi_rep2 and NMRI_2ndInf_7dpi_rep2) have more pronounced spikes in this normalised data. Additionally, samples Rag_1stInf_5dpi_rep1 and Rag_1stInf_5dpi_rep2 also have one additional peak in the lower range of the distribution, as do some other samples, but less pronounced.

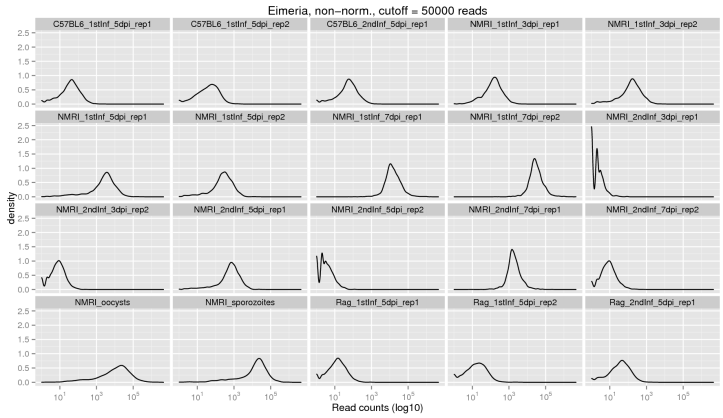

Figure 10: Density distribution of raw transcript counts (log10) for parasite samples with a cutoff of 50000 (see Methods for details). Upon visual inspection distributions in most samples appear smooth. Non-negative binomial samples in Figure 8 and 9 (NMRI_2ndInf_3dpi_rep1, NMRI_2ndInf_3dpi_rep2, NMRI_2ndInf_5dpi_rep2 and NMRI_2ndInf_7dpi_rep2 have more than one density peak also when lowly expressed genes are removed with a high threshold (50000). Samples Rag_1stInf_5dpi_rep1 and Rag_1stInf_5dpi_rep2, however, seem smoother with only one density peak with this threshold.

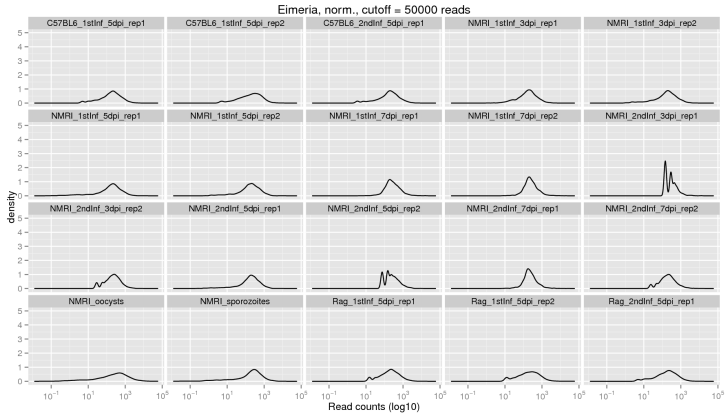

Figure 11: Density distribution of normalised transcript counts (log10) for parasite samples with a cutoff of 50000 (see Methods for details). Upon visual inspection distributions in most samples appear smooth. Non-negative binomial samples are pronounced bimodal in this normalised data (NMRI_2ndInf_3dpi_rep1, NMRI_2ndInf_3dpi_rep2, NMRI_2ndInf_5dpi_rep2 and NMRI_2ndInf_7dpi_rep2). Samples Rag_1stInf_5dpi_rep1 and Rag_1stInf_5dpi_rep2 are also bimodal, however less pronounced.

4 Conclusions

Mouse
The assumption of a negative binomial distribution underlies most differential gene expression analyses available. Similar bimodal expression patterns as in our mouse dataset have been reported (*e.g.* ....) and suggested to be a result of a group of non-coding RNAs being sequenced. Since we are not interested in such RNA species in this project, different cutoff values were tested to find the lowest possible threshold for removing transcripts and gain a negative binomial distribution. By visual inspection (again: should we formalise/quantify this somehow?) we chose 5000 as the cutoff for further analysis.

Parasite
The parasite data follows the expected distribution in most samples (although we have not formally tested it - check mean-var relationship to do that?). Bimodal or multipeak samples do not show unimodal distributions even with high cutoffs, *i.e.* many transcripts required to map to one gene for the reads to be included. The four most obvious spiky samples are also the samples with fewest reads sequenced overall, which can explain the uneven distribution (make table to refer to?). We will exclude these four samples (NMRI_2ndInf_3dpi_rep1, NMRI_2ndInf_3dpi_rep2, NMRI_2ndInf_5dpi_rep2 and NMRI_2ndInf_7dpi_rep2) from all further analysis. It is not clear why there is such low coverage in these samples. Possibilities include poor infection success because of too low oocyst injection, mouse resistance due to *e.g.* unknown previous infections or other unknown variability in the experiment. Because of this insecurity, these samples are also exculded from the mouse data.

The fact that the density peak occurs at similar read counts after normalisation in mouse and parasite data (respectively) is encouraging and indicates that normalisation is successful in both datasets.