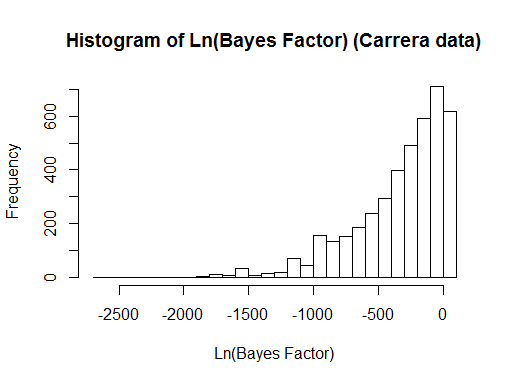
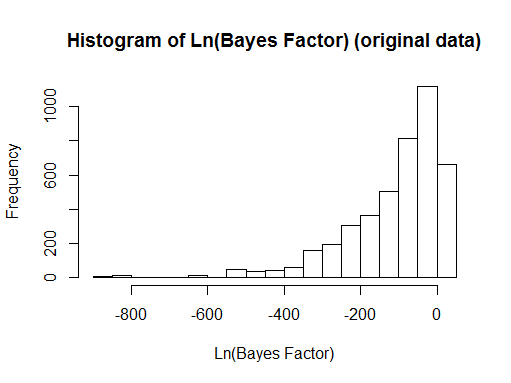
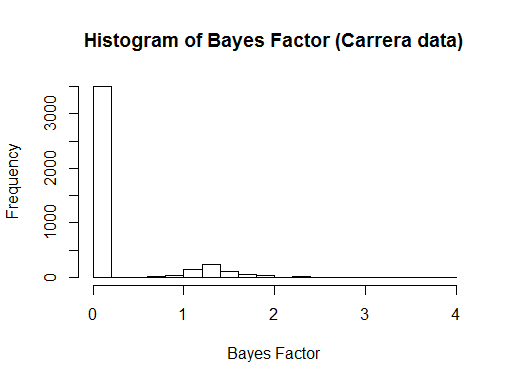
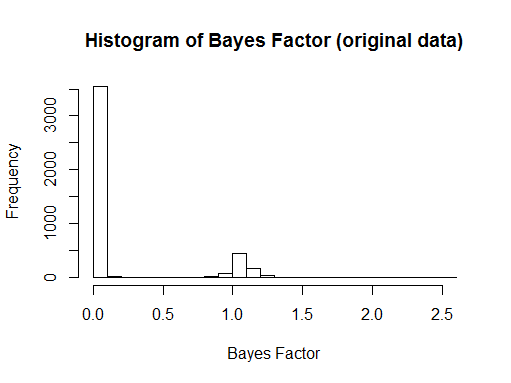
# Scores

In this document we’ll talk about three different measures (scores) of a given gene or operon. For all three of these, all genes in the same operon will always receive the same score – scores are assigned by operon. We’ll briefly describe each of them and give some statistics to give you a general idea of the distributions of these – what is a high and what is a low score. In all three cases, higher values (or less negative values) indicate more likely to be always-inactive.

## Bayes Factor

This is a measurement of the not-changing-state-ness of an operon; higher scores indicate a greater likelihood that the operon is not-changing-state.

In all of these histograms (and everywhere else in this document), each gene, not each operon, gets a data point.



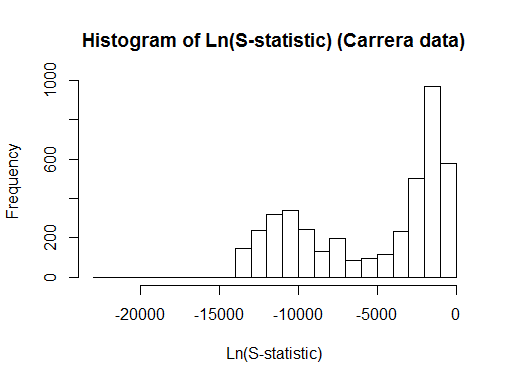
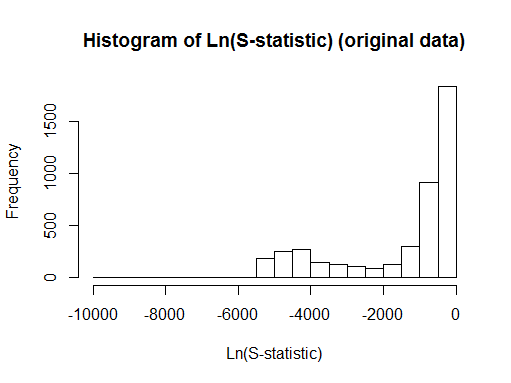
In both plots, there’s a group of genes with Bayes Factor near or over 1.0, and a group of genes with Bayes Factor much less than 1.0. A Bayes Factor over 1.0 indicates statistical evidence in favor of the 1-component model. The genes over 1.0 on the Bayes Factor histograms are the ones over 0 on the Ln(Bayes Factor) histograms. Also, the three genes in the araBAD operon are some, but not all, of the genes below -800 on the natural-log histogram on the original data; but they are towards the top of the natural-log histogram on the Carrera data, with a value of -170.2. In fact, on the Carrera data, araBAD seems a good candidate for being always-inactive, based on its scores. This is discussed further in the companion document “Always-inactive candidates.docx”.

In the original data, the highest observed value is 2.58, although it is difficult to see on the histogram. In the Carrera data, the three highest values, 67.11 and a two-gene operon with 20.53, were omitted from the Bayes Factor histogram so that they didn’t dominate the trend. The next highest value is 3.96.

Although the Ln(Bayes Factor) histograms show similarly-shaped distributions, the Carrera one appears to be scaled by a little over 2x when compared to the original one. This matches the fact that the Carrera dataset has a little over 2x the experiments (observations).

## S-statistic

This is a measure of how consistently low the aijs for an operon are. Functionally, it mostly serves to distinguish always-active from always-inactive operons after they have been identified by their Bayes Factor. A higher (less negative) S-statistic indicates the operon is more likely to be always-inactive.

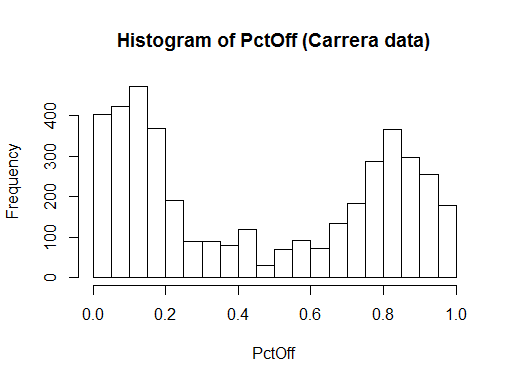
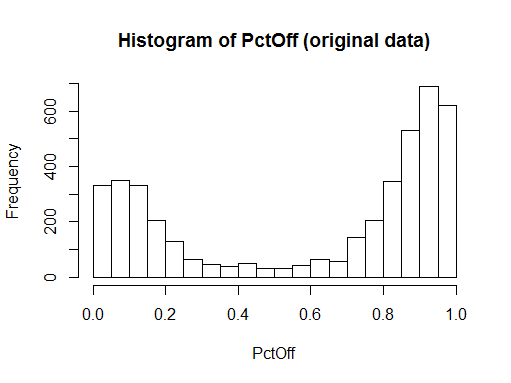


It is difficult to see on these plots, but in the original data there is a single gene at -9791.9; the next lowest gene is at -5448.1. In the Carrera data, the lowest gene is at -22448.8, followed by -18540.0.

As for Ln(Bayes Factor), the Carrera histogram for Ln(S-statistic) appears to be scaled by a little over 2x when compared to the original one. This, again, matches the fact that the Carrera dataset has a little over 2x the experiments (observations).

## PctOff

As an alternative to the S-statistic, this simply measures what percentage of the aijs for a given operon are under 0.5. Of course, a higher percentage indicates that the operon is more likely to be always-inactive (as opposed to always-active). Shown are histograms for both the original and the Carrera data.



For both datasets, most genes have either most of their aijs over 0.5, or most of their aijs under 0.5. That is, most genes are not evenly split, with values of PctOff near 0.5. In the original data we would definitely expect this, since something like 70% of our experiments are in “Complete” media (I don’t remember the exact number right now), and we expect similar gene expression data for all of those experiments. I’m not sure whether a similar pattern of similarity holds for the Carrera dataset, but I wouldn’t be surprised.

Perhaps notable is that the Carrera data has more genes in the <0.2 segment (almost-always-active) than in the >0.8 segment (almost-always-inactive), which is the reverse of the original data. Also, it’s possible that the original data has slightly more extreme values than the Carrera data here.