In this document, I have compared different essential gene predictors. These include BioTraDIS [Barquist et al.,] (blue), log fold changes from Monte Carlo method [Turner et al.,] (green), P-values from running DESeq and Monte Carlo sampling (red), the ratio between the length of the longest uninterruped segment of the gene and the length of the gene (orange), and the average distance between insertion sites. The last two methods are selected from [Freed et al.,] Figure S3. To calculate one tailed P-values from two tailed P-values obtained from DESeq, I have divided the P-values by two, and if their corresponding logFC value was less than zero, pval = 1 - pval. The set of essential genes are obtained by comparing the genes in each strain to the essential genes in E.coli K-12 from ecogene study [Zhou and Rudd,]. The results are shown in Figure 1

In Figure 2, I have compared the logodds value obtained from BioTraDIS to both logFC and P-values from Monte Carlo method for CS17 strain.

In Figure 3, I have plotted the distribution of each method. The essential and non-essential genes are well separated in Monte Carlo logFC, largest uninterrupted fraction, and mean distance between inserts. However, it is hard to distinguish between non-essential genes and beneficial losses in all these methods.

So far, Monte Carlo with logFC values is the best method. Interestingly, even though the mean distance between inserts and largest uninterrupted fraction methods are very simple, they work pretty well.

References

Barquist et al., . Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boinett, C. J., Page, A. J., Langridge, G. C., Quail, M. A., Keane, J. A., and Parkhill, J. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. 32(7):1109–1111.

Freed et al., . Freed, N. E., Bumann, D., and Silander, O. K. Combining shigella tn-seq data with gold-standard e. coli gene deletion data suggests rare transitions between essential and non-essential gene functionality. 16(1):203.

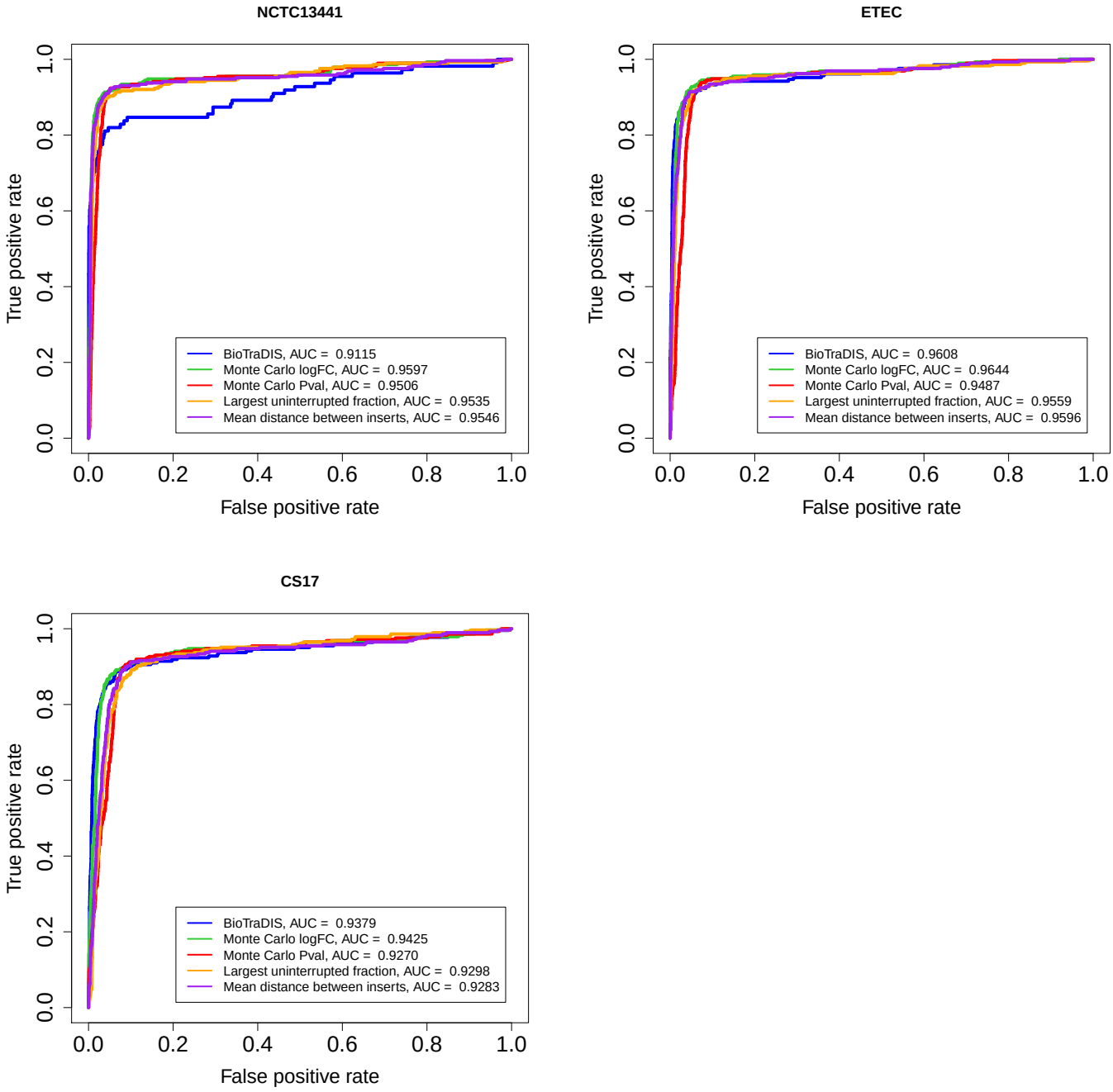

Figure 1. The comparison of different methods for calling essential genes in three E.coli strains.

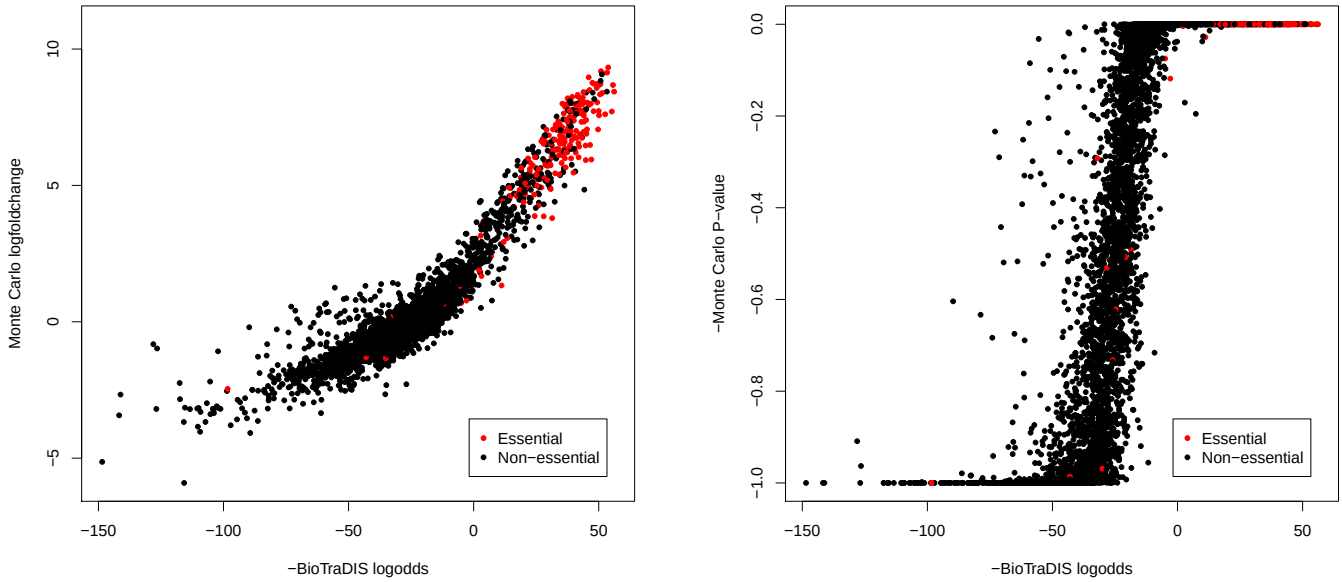

Figure 2. The figure shows the results of BioTraDIS vs. logFC from Monte Carlo method at left and BioTraDIS vs. DESeq P-values at right.

Turner et al., . Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L., and Whiteley, M. Essential genome of pseudomonas aeruginosa in cystic fibrosis sputum. 112(13):4110–4115.

Zhou and Rudd, . Zhou, J. and Rudd, K. E. EcoGene 3.0. 41:D613–D624.

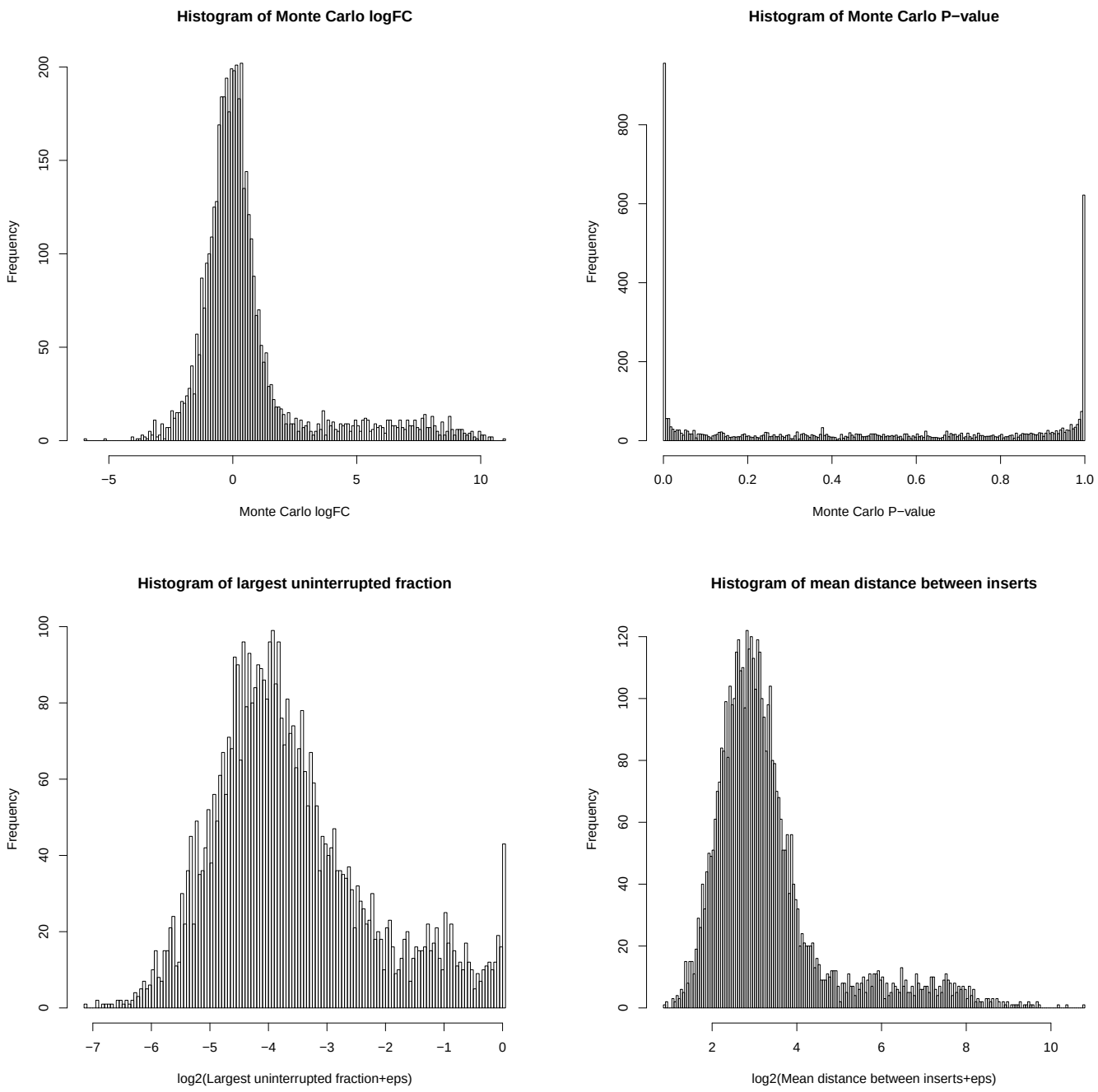

Figure 3. The figure shows the distribution of Monte Carlo logFC, Monte Carlo P-value, log2(largest uninterrupted fraction), and log2(mean distance between inserts)