# Methylated Fraction Estimation

**Estimation Method Description.**

The IPD measurements are normalized by the median value per subread. Then, for each site in the reference genome, we collect the log-transformed normalized IPDs $x_1$, …, $x_{N.}$ These values are capped as described in the kineticsTools documentation.

We fit a two-component normal mixture to this collection of normalized IPD values:

$$\sum_i \log(\,(1\ -\mathrm{p})\ \mathrm{N}(x_i\ ;\ \mu_0\,,\ \mu_0^2) + \mathrm{p}\ \mathrm{N}(x_{i\ ;}\,\mu_1\,,\ \mu_1^2)\ )$$

For each mixture component, assume that the variance equals the square of the mean, as in an exponential distribution.

The component means μ0 and μ1 are obtained from the kinetic model that is included with the kineticsTools software.

Maximum likelihood yields an estimate of p, the methylated fraction. This estimate appears as 'frac' in the kinetics.h5 output.

Finally, we use simple case resampling to obtain a percentile confidence interval for the methylated fraction. The upper and lower bounds of the 95% confidence interval appear as 'fracLow' and 'fracUp' in the kinetics.h5 output.

**Estimate of Accuracy.**

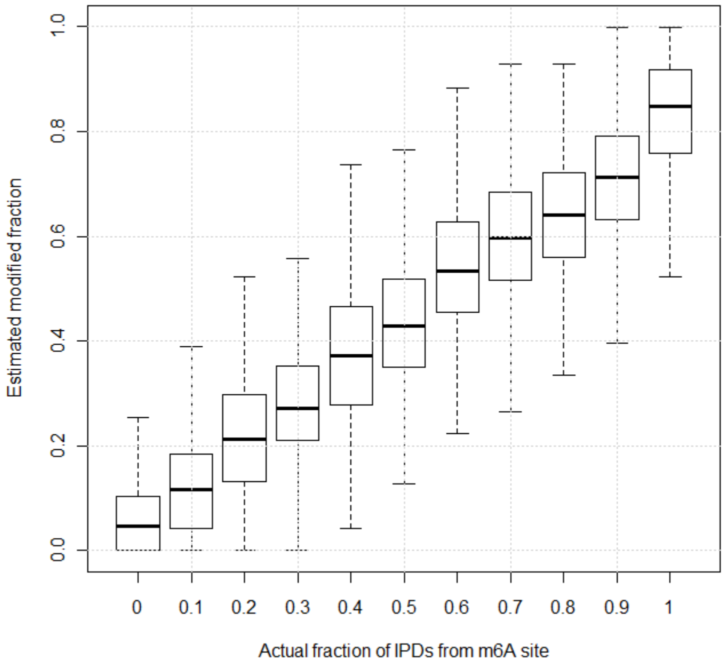
To assess the accuracy of this method, we collected IPDs from two sites in a native lambda dataset (collected internally).

One site was known to be approximately 100% methylated with m6A. The other site is believed to be completely unmethylated.

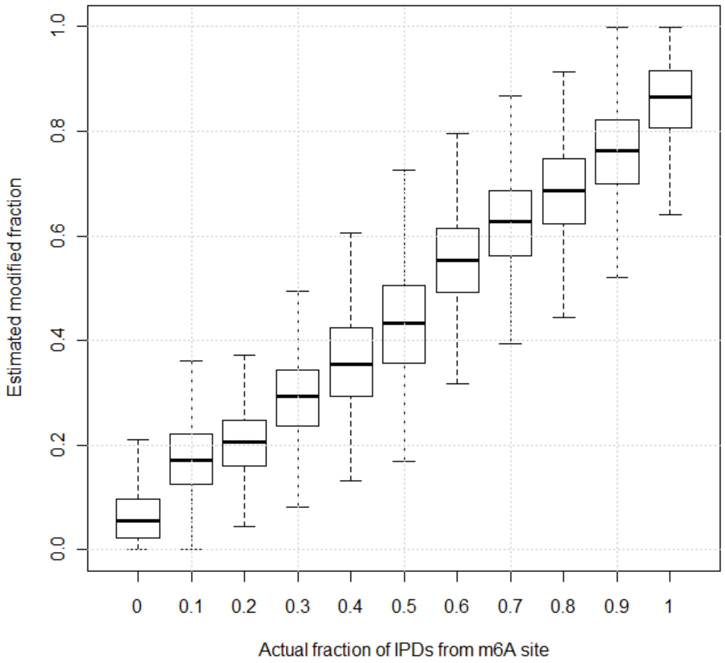
Using these two sites, we created artificial mixtures and compared the estimated methylated fraction with the known proportion of m6A values.

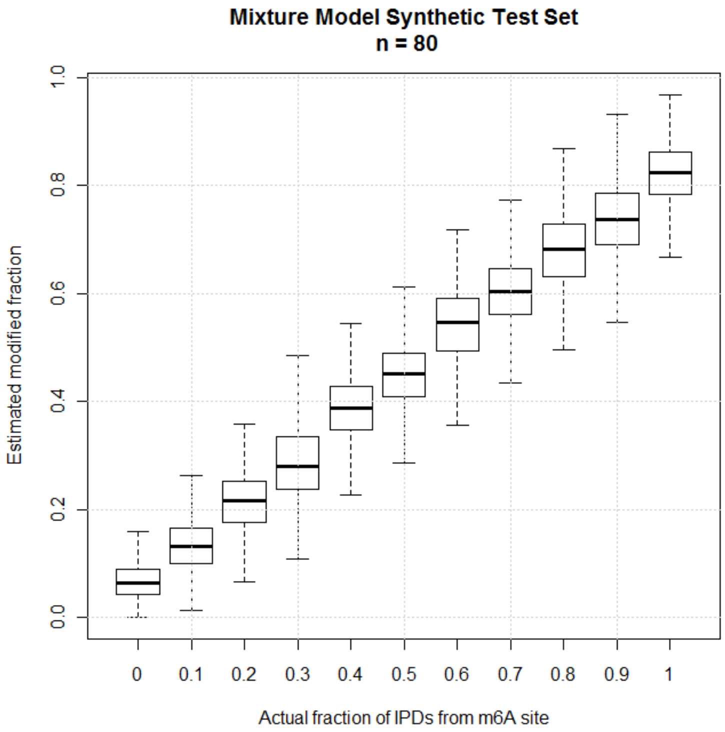
This was repeated at a range of coverage levels. The confidence intervals are expected to tighten as coverage increases:

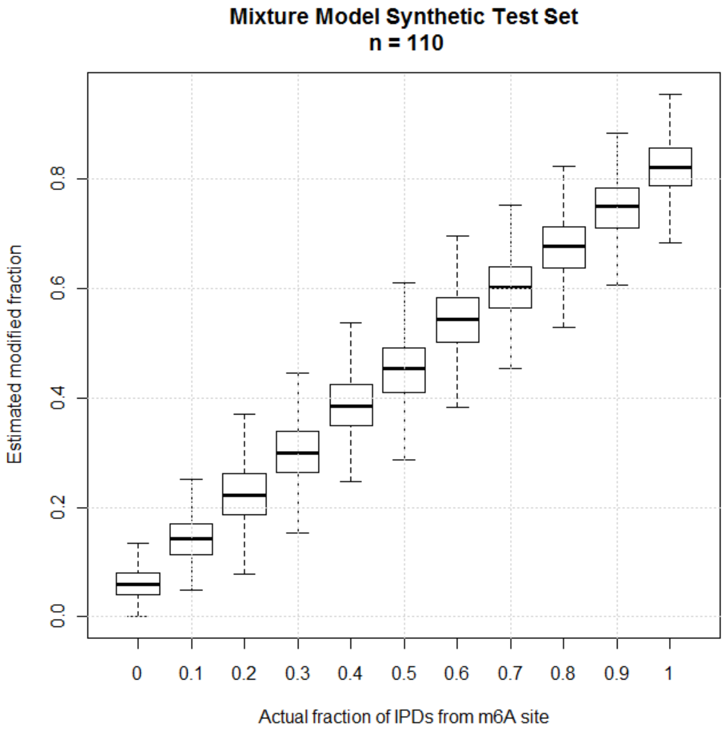
Mixture Model Synthetic Test Set
n = 25

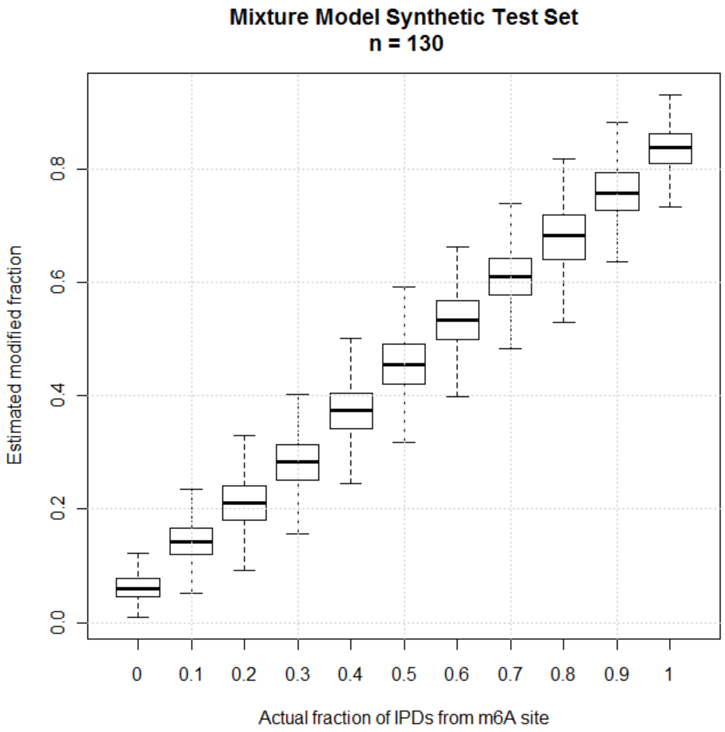


Mixture Model Synthetic Test Set
n = 50



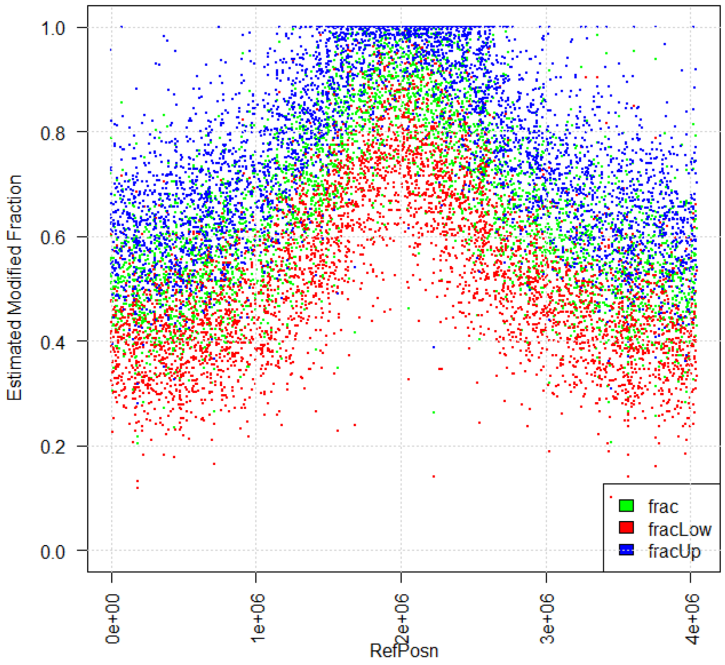

Mixture Model Synthetic Test Set
n = 80
Estimated modified fraction
0.0
0.2
0.4
0.6
0.8
1.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Actual fraction of IPDs from m6A site


Mixture Model Synthetic Test Set
n = 110
Estimated modified fraction
0.0
0.2
0.4
0.6
0.8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Actual fraction of IPDs from m6A site

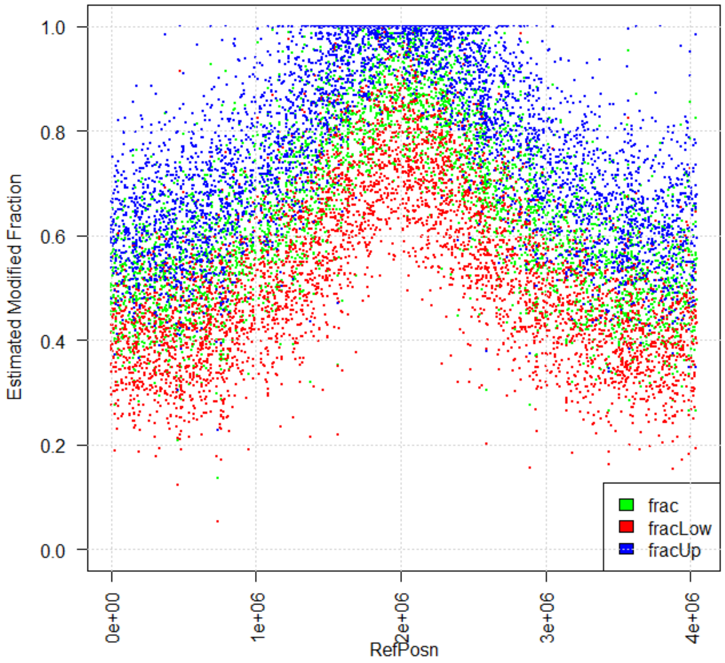


**Additional internally collected data:**

Plot the estimated methylated fraction as a function of the position on the reference genome for a Caulobacter EPD sample:



**Caulobacter EPD (51568) - Fwd**



**Caulobacter EPD (51568) - Rev**