(SUPPLEMENTARY)
A NEW APPROACH TO BIAS CORRECTION IN HIGH-THROUGHPUT SEQUENCING DATA

Daniel C. Jones, Walter L. Ruzzo, Xinxia Peng, Michael G. Katze

January 5, 2012

1 Trimming Reads

Observing the nonuniform distribution of nucleotide frequencies surrounding the 5' end of reads, a natural step to take would be to trim the 5' end before mapping, in the hope that simply removing the portion of the read in which the bias occurs will also remove the bias. Figure 1 demonstrates that this is not the case. Trimming the initial heptamer in the Mortazavi data set does nothing to reduce the bias, and simply shifts the plot by seven positions. This indicates that the issue is *sampling bias*, rather than a bias in base calling.

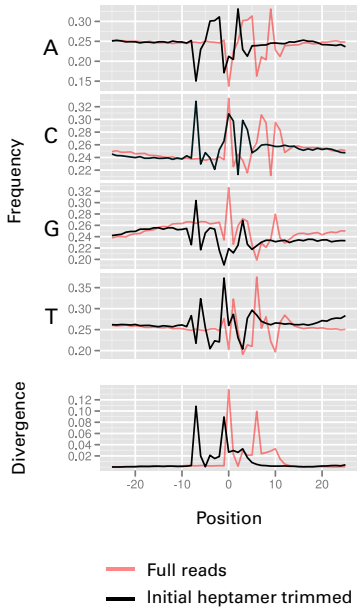

Figure 1: Nucleotide frequencies observed before (plotted in red) and after (plotted in black) trimming the initial heptamer.

2 Sensitivity of Parameters

The performance of our method is dependent on two parameters: the standard deviation at which background sequences are sampled, and the degree to which model complexity in penalized. We have also introduced parameters limiting the number of parents a node may have ($p_{\max}$) as well as the distance that an edge may traverse ($d_{\max}$), but these exist only to control the amount of CPU time used and have a much simpler interpretation: bigger is better, but slower.

Here we retrain the model on 50,000 reads from the Mortazavi data set, varying these first two parameters to assess the sensitivity of the observed results to their values.

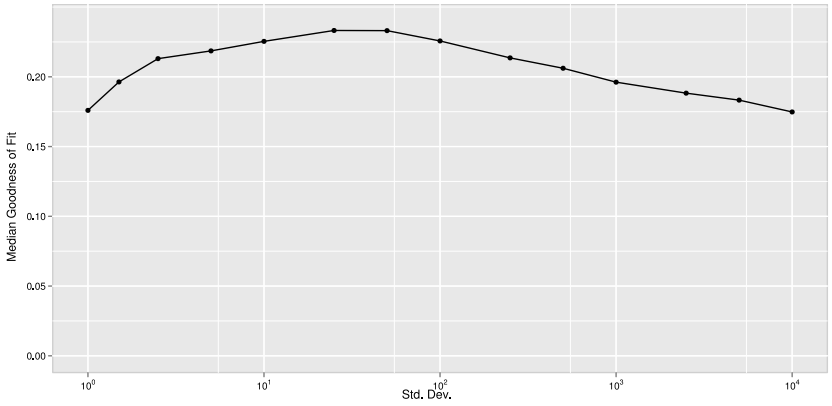

Figure 2: Sensitivity of the parameter controling the standard deviation at which background sequences are sampled, as evaluated with McFadden's R^2 goodness-of-fit statistic.

Figure 2 shows the median McFaddens R^2 goodness-of-fit statastic across 1000 test exons (as described in Section 3.2 of the main paper) versus the standard deviation used to draw background samples varied from 1 to 10,000. Apparent from this plot is that, while the optimal choice lies somewhere between 10 and 100, the model performs competitively for any reasonable value.

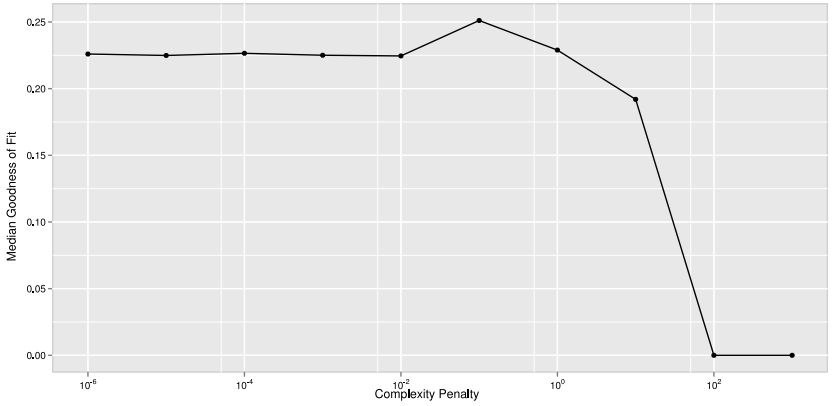

Figure 3: Sensitivity over adjustments to the complexity penalty. The Y axis is as in Figure 2.

Next we multiplied the complexity penalty term of the BIC by a constants varying from 10^{-6} to 10^3. (Recall that the BIC is $2\ell - m \log n$ where ℓ is the log-likelihood, m is the number of parameters needed to specify the model, and n is the number of training examples. Here we compute $2\ell - cm \log n$, varying c.)

It is clear from Figure 3, that a very severe complexity penalty will result in a poor model. Here, for constants larger than 100, an empty model is trained, resulting in a median R^2 of 0. Conversely, if this constant is very small, a overly-dense model will be trained. In this data, for constants less than 0.01, the maximally dense model is trained (given the restraints on in-degree and edge distance). This result is a sub-optimal but entirely adequate solution.

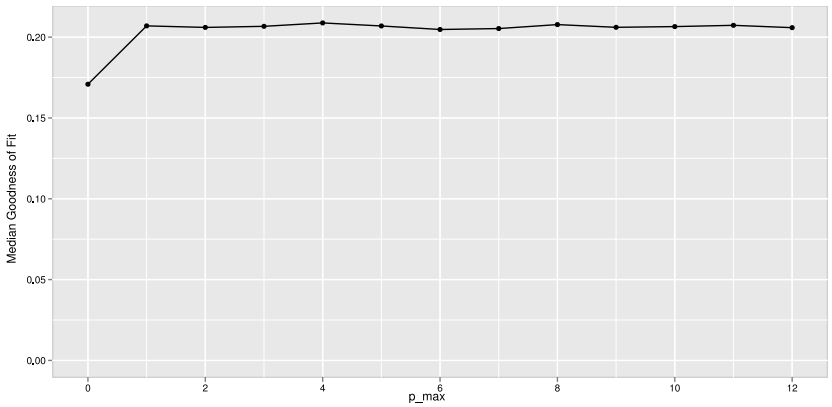

Figure 4: Sensitivity over values of the p_{max} parameter, controlling the maximum in-degree of any node in the directed graph. The Y axis is as in Figure 2.

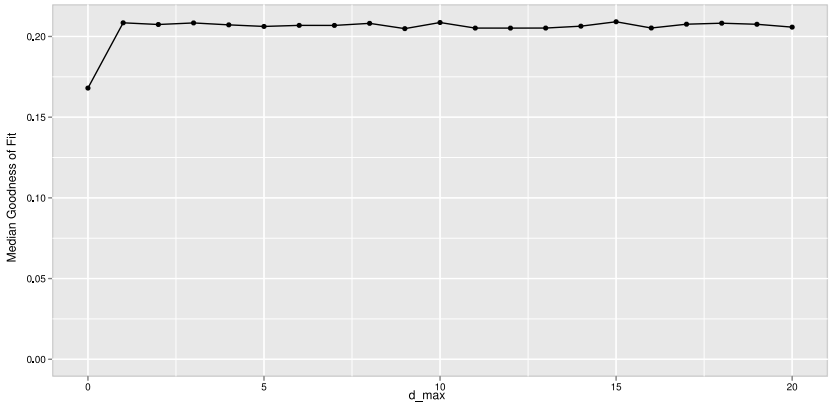

Figure 5: Sensitivity over values of the d_{max} parameter, controlling the maximum distance any edge in the directed graph may traverse. The Y axis is as in Figure 2.

Figures 4 and 5 suggest that most dependencies are between adjacent nucleotides, as increasing the maximum in-degree ($p_{\max}$), or maximum edge distance ($d_{\max}$) has little effect for values of at least 1.

3 Total Numbers of Reads Used to Train Each Method

Experiment	Total	7mer	GLM/MART	BN
Wetterbom	3.9×10^7	2.3×10^7	5.9×10^6	1×10^5
Katze	4.8×10^7	2.0×10^7	3.1×10^6	1×10^5
Bullard	1.2×10^7	3.1×10^6	1.8×10^5	1×10^5
Mortazavi	3.9×10^6	1.9×10^6	3.8×10^5	1×10^5
Trapnell	3.4×10^7	1.6×10^7	5.9×10^6	1×10^5

Table 1: Note that in the Trapnell data set (the only paired-end set considered) we count only the first mate of each read.

4 Additional Kullback-Leibler Divergence Plots

Figure 6 plots the adjusted and unadjusted KL divergence plots, supplementing those in the results section of the main paper.

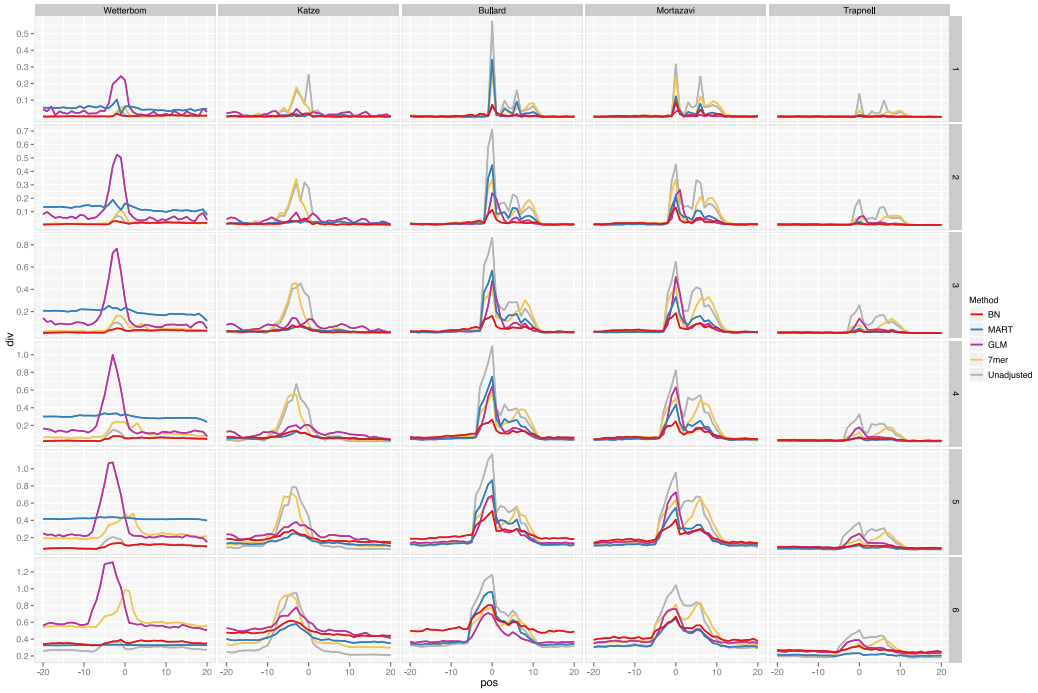

Figure 6

5 Sequence bias in data from Au, et al., 2010

Figure 7 plots sequence bias and KL divergence in the data published by Au, et al. [1] and examined by us in Section 3.3.

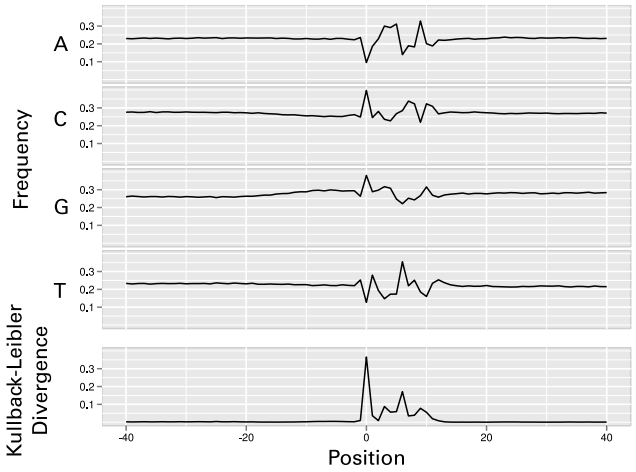

Figure 7

6 Runtime-Accuracy Trade-off

Training our model on more reads results in more accurate estimation of bias, but a longer training time. Here we investigate that trade-off by training our model on progressively more reads from the Mortazavi data set. For each subset we train our method, evaluate the median pseudo-coefficient of determination (R^2) over exons selected for our test set, and record the training time required on one core of a 3Ghz Intel Xeon processor. These results are plotted in Figure 8.

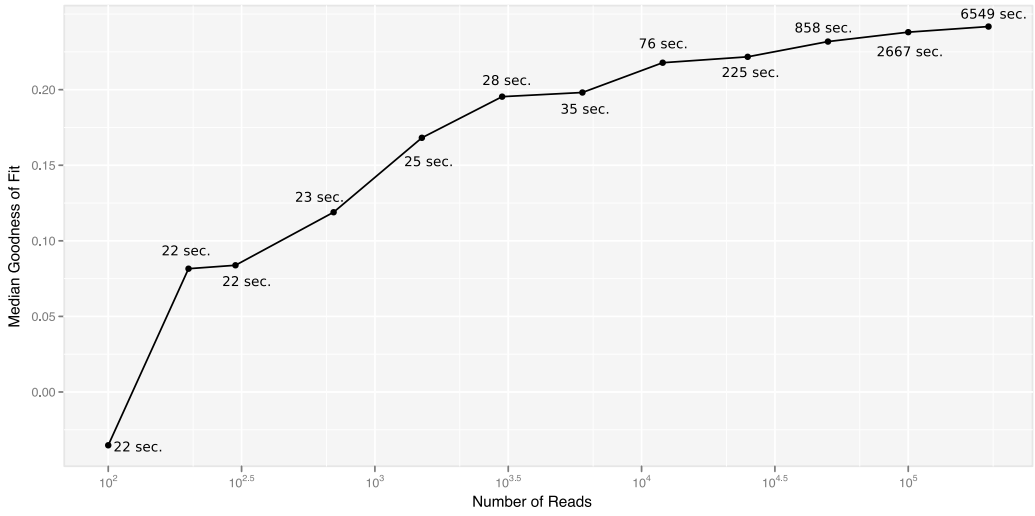

Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.

7 ChIP-Seq

Though the MART model [6] performed very well in several cases, ours offers the advantage that no gene annotations are required for training. In RNA-Seq this is useful in applications of de-novo gene discovery, but it also allows our method to be applied to ChIP-Seq and other short read data. Here we examine one publicly available ChIP-Seq data set from Cao, et. al. [3]. Specifically, we used one run, with Sequence Read Archive accession number SRR034831, containing 3,873,192 reads. These were mapped to the reference genome using Bowtie [5], resulting in 1,382,867 uniquely mapped reads, which were used for this analysis.

Measuring nucleotide frequencies, we find the bias to be significantly less than that observed the RNA-Seq data sets we tested, but the data was by no means unbiased, as shown in Figure 9.

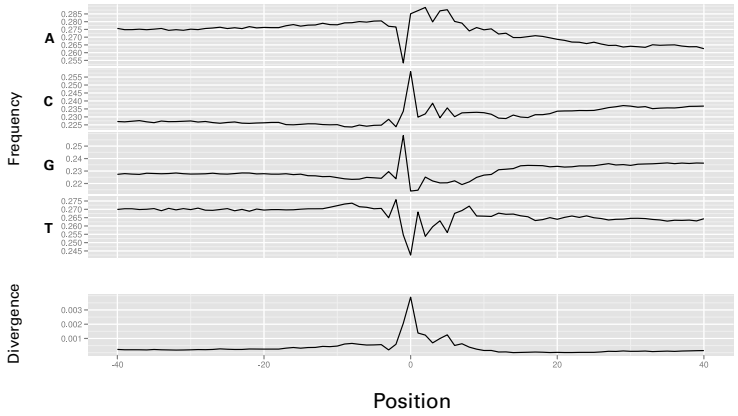

Figure 9: Nucleotide frequencies and KL divergence for the Cao data set.

In the analysis of the RNA-Seq data, we used the assumption of continuous transcription across annotated exons to evaluate the efficacy of the models, as well as to train the GLM and MART models. Evaluating bias correction on ChIP-Seq data necessitates dropping the Poisson regression test, as well as the GLM and MART models from our comparison, which can not be applied to such data.

We did however repeat our analysis using the Kullback-Leibler divergence. We used several variations of the method described by Hansen, et. al., [4]. "7mer" estimated the initial heptamer, "Avg-7mer" averages the initial two heptamers, and "4mer" averages the initial two 4mers.

We trained each method using reads from chromosomes 1–8. The remaining chromosomes were segmented into 500nt bins, and the KL divergence was sampled from the 50,000 segments with the highest read counts.

Figure 10, we plot directly the positional KL divergence, computed in the same manner described in the Results section of the paper. We see that our method is effective at reducing the bias in this ChIP-Seq data.

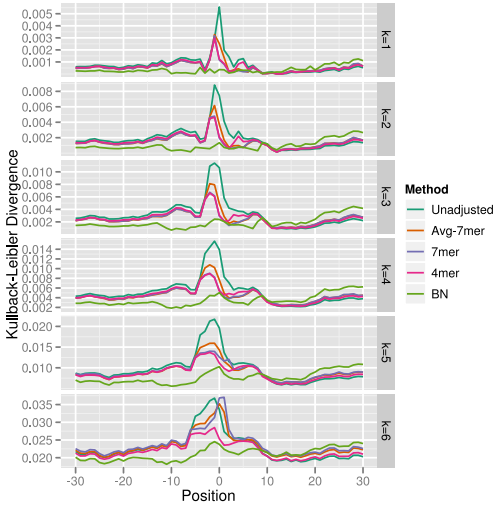

Figure 10: KL divergence for the Cao data set, after adjusting read counts for bias.

8 Bias in Amplification-Free Sequencing

To evaluate the extent to which the observed bias is caused by PCR amplification in the library preparation stage, we analyzed data from the FRT-Seq protocol developed by Mamanova, et. al. [7]. FRT-Seq avoids the amplification step during library preparation with reverse transcription occuring on the flowcell surface. We obtained data generated by Mamanova, et. al., from the European Nucleotide Archive with accession code ERR007689, and mapped them to the hg19 assembly of the human genome with Bowtie [5]. The nucleotide frequencies and divergence are plotted in Figure 11.

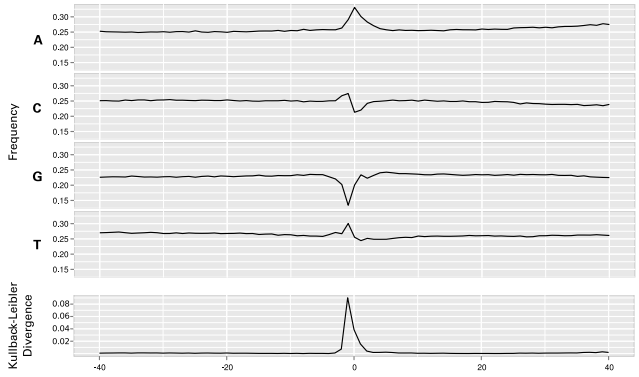

Figure 11: Nucleotide frequencies and KL divergence of the Mamanova data set.

It is clear from this analysis that the FRT-Seq method, as implemented to generate this data, is not without bias, yet there is significant divergence at only a small number of positions near the read start, including the position before it. When our method is trained on this data we obtain the suitably sparse model pictured in Figure 12.

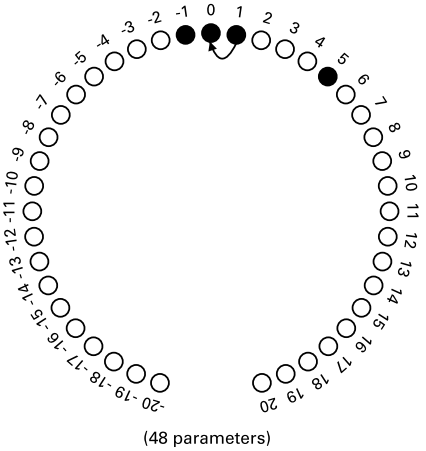

Figure 12: The graphical model learned by training our method on the Mamanova data set.

9 Variability Between Replicates

Measuring differential expression or isoform switching is a primary application of RNA-Seq. If the bias were inconsistent between replicates, it would call into question the accuracy of such

tests. In our experiments, we have observed the bias to be mostly, but not entirely consistent between replicates. In Figure 13 we plot the frequencies from the four runs in the Trapnell data set.

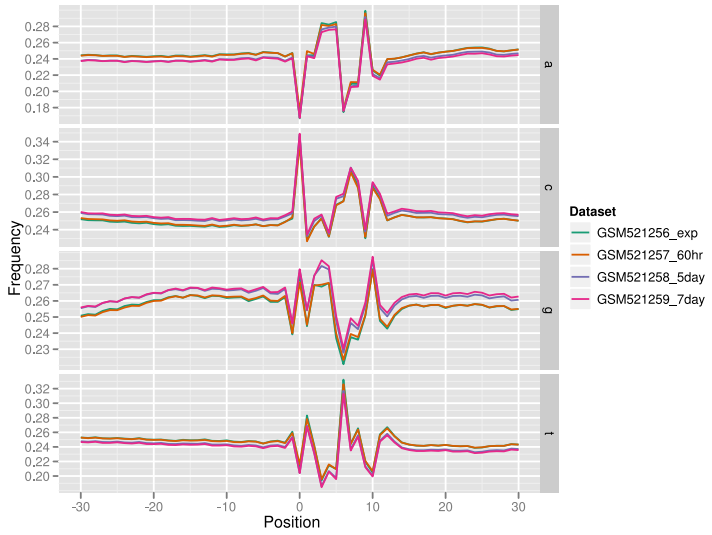

Figure 13: Nucleotide frequencies of replicates in the Trapnell data set.

Though we do not know if the variability is always minor, assuming this is so, there remains the risk that the sequence bias would greatly effect the depth to which a locus is sequenced, and thus the statistical significance of differential expression tests. This would bias the discovery of differentially expressed genes, but not the test itself. Additionally, as sequencing bias appears to be at least partly related to sequence context, alternative splicing potentially will change the bias at positions somewhat distant from the splice site itself, complicating any analysis of differential isoform usage.

10 KL Divergence and Parameter Estimation

10.1 Overview

In this section and the next, we present some theoretical analysis quantifying the accuracy of parameter estimation and the likelihood of falsely learning a biased model from unbiased data. These results follow mostly from work originally done by Birch [2], but are presented here for completeness, and in a manner more directly addressing the the question at hand.

Suppose P and Q are probability distributions defined on a discrete sample space. E.g., in the context of this paper, think of them as the probabilities of DNA k-mers for some fixed k. The *Kullback-Leibler divergence*, also known as *relative entropy*, of Q with respect to P is defined as

$$H(Q||P) = \sum_i q_i \ln \frac{q_i}{p_i}$$

where q_i (p_i) is the probability of observing the i^{th} event according to the distribution Q (resp., P), and the summation is taken over all events in the sample space (e.g., all k-mers). In some sense, this is a measure of the dissimilarity between the distributions: if $p_i \approx q_i$ everywhere,

their log ratios will be near zero and H will be small; as q_i and p_i diverge, their log ratios will deviate from zero and H will increase.

For a more quantitative and (perhaps) intuitive interpretation, consider the following hypothesis-testing scenario. Given m independent samples from either P (arbitrarily called the "null") or from Q (the "alternative"), how large should m be to confidently choose between these cases? A natural approach is to use a likelihood ratio test: letting Y_i be the number of times event i is observed in m trials ($\sum Y_i = m$), the likelihood of the sequence of observations under model P is

$$\prod_i p_i^{Y_i}$$

and similarly for Q. So, the logarithm of the likelihood ratios is

$$\text{LLR} = \ln \frac{\prod_i q_i^{Y_i}}{\prod_i p_i^{Y_i}} = \sum_i Y_i \ln \frac{q_i}{p_i} = m \sum_i \frac{Y_i}{m} \ln \frac{q_i}{p_i}$$

If the sample is drawn from the alternative distribution Q, then the expectation of $\frac{Y_i}{m}$ is q_i, so the expected value of the LLR is

$$m \sum_i q_i \ln \frac{q_i}{p_i} = mH(Q||P)$$

That is, the KL divergence is exactly the expected per-sample contribution to the log-likelihood ratio. So, assuming the null hypothesis is false, in order for it to be rejected with say, $1000 : 1$ odds, one should choose m to be inversely proportional to $H(Q||P)$:

$$mH(Q||P) \geq \ln 1000$$
$$m \geq \frac{\ln 1000}{H(Q||P)}$$

As a concrete example, all of the RNA-Seq data sets examined in this paper show a KL divergence between the 1st position of reads versus the transcriptomic background nucleotide distribution of 0.05 or greater. Thus, one can confidently reject the hypothesis that reads are beeing uniformly sampled across the transcriptome by examining only a few hundred randomly selected reads (i.e., $\ln 1000/0.05 \approx 140$), a surprisingly small number, given the supposedly "random sampling" accomplished by RNA-Seq.

10.2 Accuracy of Multinomial Parameter Estimation

Continuing the notation above, suppose P as an unknown distribution with parameters $p_1, \ldots, p_r$, $\sum p_i = 1$ where r is the number of points in the sample space (e.g. $r = 4^k$ in the case of k-mers). Given a random sample $X_1, X_2, \ldots, X_r$ of size $n = \sum_i X_i$ from P, it is well known that the maximum likelihood estimators for the parameters are $q_i = \frac{X_i}{n} \approx p_i$. How good an estimate for P is this distribution Q? The estimators are unbiased:

$$E[q_i] = E\left[\frac{X_i}{n}\right] = \frac{E[X_i]}{n} = \frac{np_i}{n} = p_i$$

and the standard deviation of each estimate is proportional to $1/\sqrt{n}$, so these estimates are increasingly accurate as the sample size increases. A more quantitative assessment of the accuracy of the estimator is obtained by evaluating the KL divergence:

$$H(Q||P) = \sum_{i=1}^{r} q_i \ln \frac{q_i}{p_i} = \sum_{i=1}^{r} q_i \ln \left(1 + \frac{q_i - p_i}{p_i}\right)$$

Using the first two terms of the Taylor series for $\ln(1+x)$, this is

$$\begin{aligned} H(Q||P) &\approx \sum_{i=1}^{r} q_i \left(\frac{q_i - p_i}{p_i} - \frac{1}{2} \left(\frac{q_i - p_i}{p_i} \right)^2 \right) \\ &= \sum_{i=1}^{r} q_i \frac{q_i - p_i}{p_i} - \frac{q_i}{2p_i} \frac{(q_i - p_i)^2}{p_i} \end{aligned}$$

Since $\sum_{i=1}^{r} q_i = \sum_{i=1}^{r} p_i = 1$, $\sum_{i=1}^{r} p_i \frac{q_i - p_i}{p_i} = 0$, so

$$\begin{aligned} H(Q||P) &\approx \sum_{i=1}^{r} q_i \frac{q_i - p_i}{p_i} - p_i \frac{q_i - p_i}{p_i} - \frac{q_i}{2p_i} \frac{(q_i - p_i)^2}{p_i} \\ &= \sum_{i=1}^{r} \frac{(q_i - p_i)^2}{p_i} \left(1 - \frac{q_i}{2p_i} \right) \\ &\approx \frac{1}{2} \sum_{i=1}^{r} \frac{(q_i - p_i)^2}{p_i} \end{aligned}$$

since $q_i \approx p_i$. Multiplying by n^2/n^2 we have,

$$\begin{aligned} H(Q||P) &\approx \frac{1}{2n} \sum_{i=1}^{r} \frac{(nq_i - np_i)^2}{np_i} \\ &= \frac{1}{2n} \sum_{i=1}^{r} \frac{(X_i - E[X_i])^2}{E[X_i]} \end{aligned}$$

The summation is the test statistic for the χ^2 goodness-of-fit test for a multinomial distribution, and as $n \to \infty$ is known to follow a χ^2 distribution with $r-1$ degrees of freedom. Finally, the expected value of such a random variable is $r-1$, hence the expected KL divergence of the MLE inferred distribution Q with respect to the true distribution P is

$$E[H(Q||P)] = \frac{r-1}{2n} \tag{1}$$

Stochastic simulations of this, shown in Figure 14, proves when P is uniform show that the above formula is a very good fit when $n > r$.

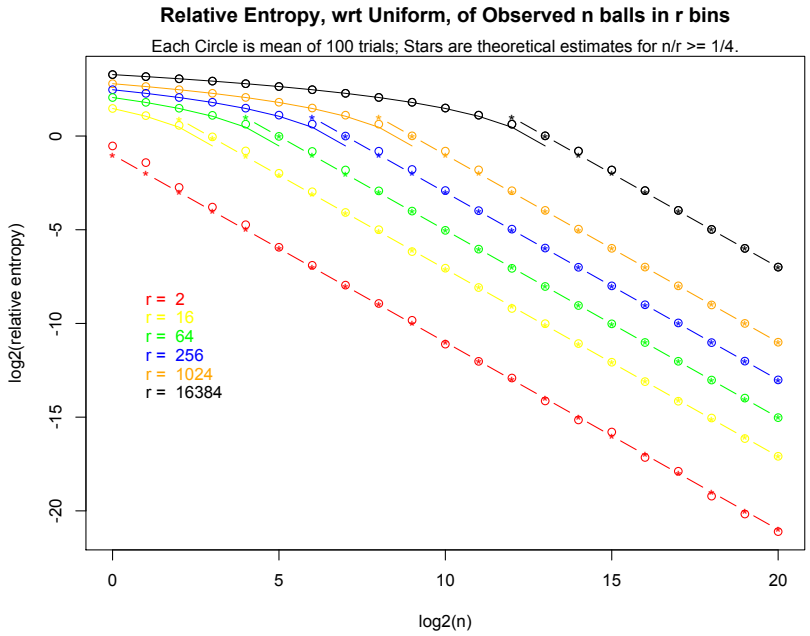

Figure 14: For each value of n plotted, the circles indicate the value of $H(Q||P)$, averaged over 100 samples of size n, where P is uniform and Q is the MLE estimator for P based on n random samples drawn from r bins. The asterisks and the straight line interpolating them are the theoretical approximation from Equation 1.

11 False Discovery of Bias in Unbiased Experiments

In the case that an experiment is unbiased, we would like an empty model to be trained, so that using the method to correct for bias will have no effect. Any non-empty model would otherwise decrease the accuracy of the quantification. To address this concern, we derive here an upper bound on the probability of a non-empty model being trained, when the data is unbiased.

The training procedure begins by considering the set of models formed by including each single nucleotide position from within the training sequences. For each single nucleotide model, the Bayesian information criterion is evaluated. If in any of these models the BIC score increases over the empty model, a non-empty model will be trained, otherwise the training procedure will halt with an empty model.

We begin by considering one single nucleotide model. Suppose the background and foreground nucleotide distributions for this single position are equal (i.e., the experiment is unbiased in this position) specified by the parameters (p_a, p_c, p_g, p_t).

The model is trained on a set of n foreground and n background sequences. Let $X = (x_a, x_c, x_g, x_t)$ give the number of times each nucleotide was observed in the foreground sample, and $Y = (y_a, y_c, y_g, y_t)$ in the background sample.

First, suppose the nucleotide distribution estimated from X and from Y both have KL divergence bounded by some number ϵ:

$$D_X = \sum_{k \in \{a,c,g,t\}} (x_k/n) \log\left(\frac{x_k/n}{p_k}\right) < \epsilon$$

and

$$D_Y = \sum_{k \in \{a,c,g,t\}} (y_k/n) \log\left(\frac{y_k/n}{p_k}\right) < \epsilon$$

Now, the objective function evaluated is the Bayesian information criterion (BIC) applied to the conditional log-likelihood (CLL).

$$\text{BIC} = 2 \cdot \text{CLL} - m \log n$$

where CLL in this single nucleotide case can be written as,

$$\text{CLL} = \sum_{k \in \{a,c,g,t\}} \left(x_k \log\left(\frac{x_k/n}{(x_k+y_k)/n}\right) + y_k \log\left(\frac{y_k/n}{(x_k+y_k)/n}\right)\right)$$

The KL divergence obtained by combining the two samples X and Y is,

$$D_{XY} = \sum_{k \in \{a,c,g,t\}} \frac{(x_k+y_k)}{2n} \log\left(\frac{(x_k+y_k)/2n}{p_k}\right)$$

The KL divergence is non-negative, and so adding $2nD_{XY}$ to the CLL, we get

$$\begin{aligned}
CLL &\leq CLL + 2nD_{XY} \\
&= \textstyle\sum_{k \in \{a,c,g,t\}} x_k \log\left(\frac{x_k/n}{(x_k+y_k)/n}\right) + y_k \log\left(\frac{y_k/n}{(x_k+y_k)/n}\right) + (x_k+y_k)\log\left(\frac{(x_k+y_k)/2n}{p_i}\right) \\
&= \textstyle\sum_{k \in \{a,c,g,t\}} \left(x_k \log\left(\frac{x_k/n}{2p_i}\right) + y_k \log\left(\frac{y_k/n}{2p_i}\right)\right) \\
&= nD_X + nD_Y + \textstyle\sum_{k \ in \{a,c,g,t\}} (x_k+y_k)\log(1/2) \\
&< 2n\epsilon + 2n\log(1/2)
\end{aligned}$$

Thus, if the KL divergence of the two samples X and Y are bounded, the conditional log-likelihood is as well.

The BIC score of an empty model is,

$$\text{BIC}_0 = 2 \sum_{k \in \{a,c,g,t\}} (x_k \log(1/2) + y_k \log(1/2)) - m \log n = 4n \log(1/2)$$

where the number of parameters needed to specify the empty model is $m = 0$.

The BIC score of the single nucleotide model, when the KL divergence of X an Y is bounded by ϵ is

$$\begin{aligned}
\text{BIC} &= 2 \cdot CLL - m \log n \\
&= 2 \cdot CLL - 6 \log n \\
&< 4n\epsilon + 4n \log(1/2) - 6 \log n
\end{aligned}$$

It follows that if

$$\begin{aligned}
4n\epsilon + 4n \log 1/2 - 6 \log n &\leq 4n \log 1/2 \\
\epsilon &\leq \frac{3}{2n} \log n
\end{aligned}$$

then

$$\text{BIC} \leq \text{BIC}_0$$

In summary, if $D_X, D_Y < \frac{3}{2n} \log n$, then BIC $\leq$ BIC$_0$ and the empty model is retained. That is, the probability of an empty model being trained is bounded below by the probability that the KL divergence of both samples is bounded above by $\frac{3}{2n} \log n$:

$$\Pr(D_X < \frac{3}{2n} \log n) \cdot \Pr(D_Y < \frac{3}{2n} \log n)$$

Since X and Y are drawn from the same distribution, D_X and D_Y are identically distributed as well, and so we can simplify this formula as,

$$(\Pr(D < \frac{3}{2n} \log n))^2$$

where D is the divergence of any sample of size n. The probability of a non-empty model being trained is then less than

$$1 - (\Pr(D < \frac{3}{2n} \log n))^2$$

In training we consider multiple nucleotides. If h nucleotides are considered, the probability of a non-empty model being trained is less than

$$1 - (\Pr(D < \frac{3}{2n} \log n))^{2h}$$

As discussed in Supplementary Section 10.2, $2nD$ is approximated by a χ^2 distribution with 3 degrees of freedom for large n, and so we can estimate this probability for various values of n. This is shown in Figure 15. From this figure we can see, for example, that training with more than 10,000 reads results in an incorrect non-empty model with probability less than 0.0004.

For small n, the upper bound on non-empty model probability is much higher. In our tests, this upper bound is not tight for a small number of reads. We trained the model on 100 sets of 100 simulated reads, drawn uniformly at random from exonic sequences and in only 3 trials was a non-empty model trained, far less than the upper bound expectation of 23 out of 100. This test was repeated using sets of 10,000 reads, and in no trial was a non-empty model trained. From these results, and the results in Supplementary Section 6, we recommend training our method with at least 10,000 reads.

Here we considered only the case of an empty model versus a non-empty model. A similar analysis can be conducted to show that if an extraneous position were added to to the initial empty model, it would be very improbable that more extraneous positions are edged were added. So that if, despite the very low probability, a non-empty model were trained from unbiased data, it would be extremely sparse and have little effect.

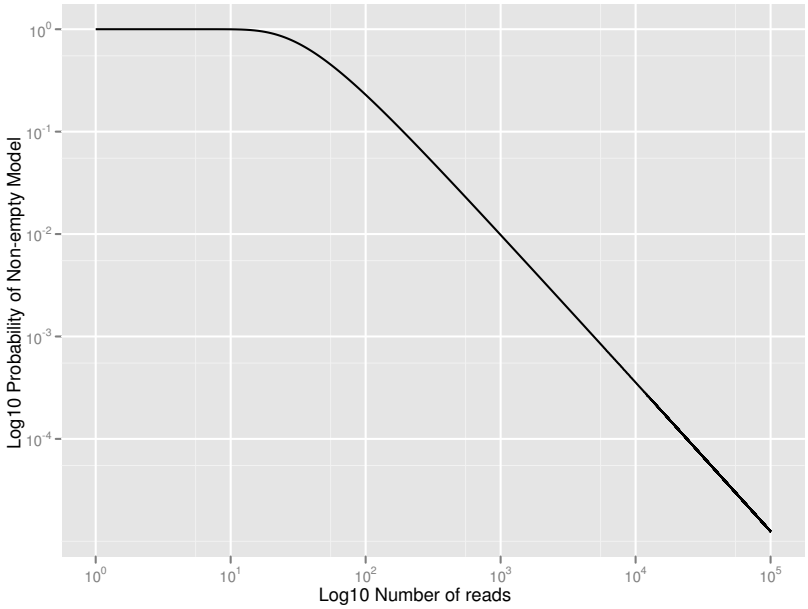

Figure 15: Plotted is the upper bound on the probability of training a non-empty model when there is no real bias in the experiment when the model is trained with various numbers of reads. We fixed the number of nucleotides being considered as $h = 41$, as was used for the evaluation in Section 3.

References

[1] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, 38(14):4570–8, August 2010.

[2] MW Birch. A new proof of the Pearson-Fisher theorem. *The Annals of Mathematical Statistics*, 1964.

[3] Yi Cao, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J Sanchez, Maura H Parker, Kyle L MacQuarrie, Jerry Davison, Martin T Morgan, Walter L Ruzzo, Robert C Gentleman, and Stephen J Tapscott. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Developmental Cell*, 18(4):662–74, April 2010.

[4] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):1–7, April 2010.

[5] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, January 2009.

[6] Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R50, January 2010.

[7] Lira Mamanova, Robert M Andrews, Keith D James, Elizabeth M Sheridan, Peter D Ellis, Cordelia F Langford, Tobias W B Ost, John E Collins, and Daniel J Turner. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature Methods*, 7(2):130–2, February 2010.