Equilibria: A Report on Comparing Resampling Methods for Robust RNA-seq Quantification Estimation

1 Motivation

Current research solves the problem of transcript quantification using expectation maximization based approaches. The problem is particularly hard because of the presence of multi-mapping reads and the variance in the estimation of the transcript at technical, biological and inference level. Problem of multi-mapping reads due to isomorphic transcripts are well studied by the community but the presence of variance at multiple levels drives the need to get a range estimate instead of point estimate per transcript in each sample without which the downstream analysis like the differential expression lose sensitivity and specificity.

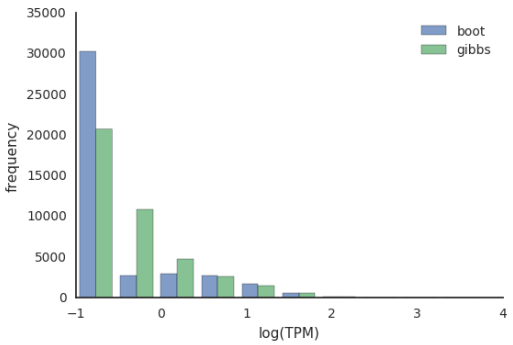
Bootstrapping and Gibbs sampling are the most prevalent resampling methods found in RNA-seq methods literature. The former is used before quantification optimization and resamples the reads from an equivalence class found by [1] while latter resamples from posterior estimates of the optimization to account for technical and inferential variance. While the current state of the art methods doesn't specify why one should be preferred over another a recent differential expression tool (sleuth) uses bootstrap for accounting variance in the analysis. This raises the question of what more can Gibbs sampling bring into the analysis. Currently, there is no comparison of these resampling methods for RNA-seq quantification estimation so in this report we have compared the two resampling methods.

2 Results

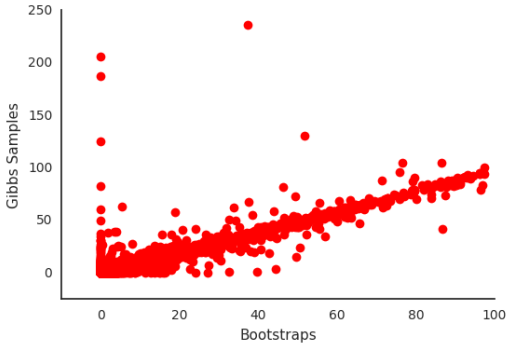
We started with one replicate and, ran bootstrap and Gibbs sampling algorithm to get 100 samples per transcript. First, the most intuitive way was to compare abundance values directly. So we took the mean abundance value of each transcript for both bootstrap and Gibbs sampling and visualize it through a histogram fig. 1a. Although they show good correlation with highly abundant transcripts but there is a significant difference in the abundance values especially around 10 TPM values. fig. 4c derives an observation that transcript abundances below 10 TPM were of special interest since they were showing totally different frequency distribution. The observation that Bootstrap and Gibbs sample behave differently becomes even more evident with the presence of transcript on the left extreme of the scatter plot. This basically signifies that some transcripts predicted as non-abundant by bootstrap have been given as large as 200 TPM by Gibbs Sample which needed further investigation.

On close observation of the transcript of interest which supposedly was 'ENST00000472311' of gene 'COX7A2' under RNA-seq experiment [2] with a COPD affected condition we found that the comparison between the posterior estimates of Bootstrap and Gibbs Sample leads to bimodal distribution as shown in fig. 3. On the left of the figure we see Gibbs Sample with bimodal frequency distribution between 0 and 200 and on the right, we have bootstrap with almost certainly all zero estimates.

To verify our findings on the real data we simulated data using polyester [3] and look for similar trends. In figure fig. 2 we plot a violin plot for a simulated data on one of the replicates. Similar to real data we find that plotting the violin plot for the coefficient of variation of posterior estimates of the transcripts on both the methods gives significantly different distributions. In particular Bootstrap tends to be more conservative in assigning reads to a transcript but with more certainty while Gibbs Sample assignment shows some variability in low abundant transcripts.

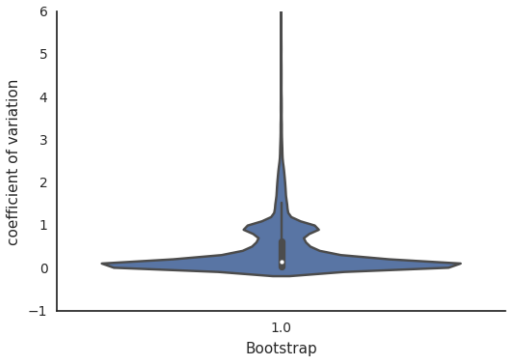


(a) Histogram of abundance values (log(TPM)) for Bootstrap and Gibbs sample in one replicate.

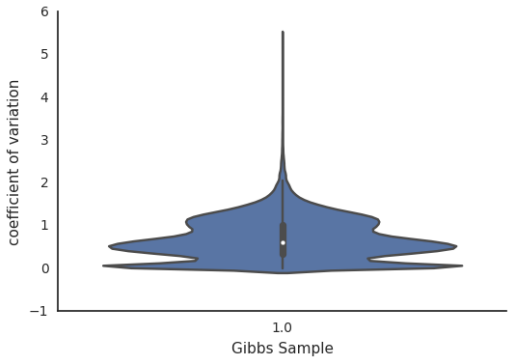


(b) Scatter plot for TPM values of Bootstrap and Gibbs Sample zoomed at 100 TPM.

Figure 1: Visualizing abundances in Real data

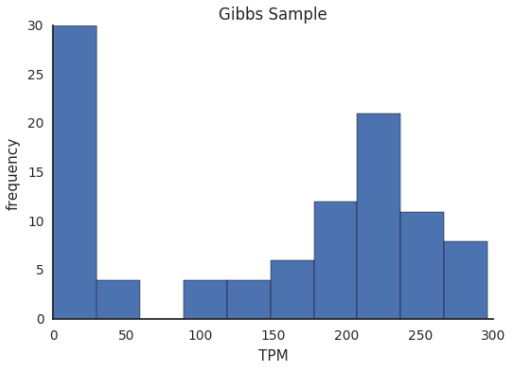


(a) Violin plot of coefficient of variation of all transcripts for Bootstrap in one one replicate.

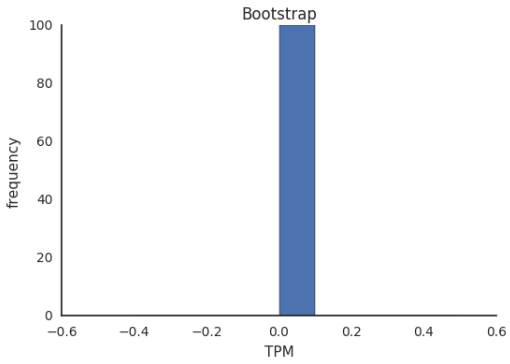


(b) Violin plot of coefficient of variation of all transcripts for Gibbs Sample in one one replicate.

Figure 2: Violin Plots on Simulated data

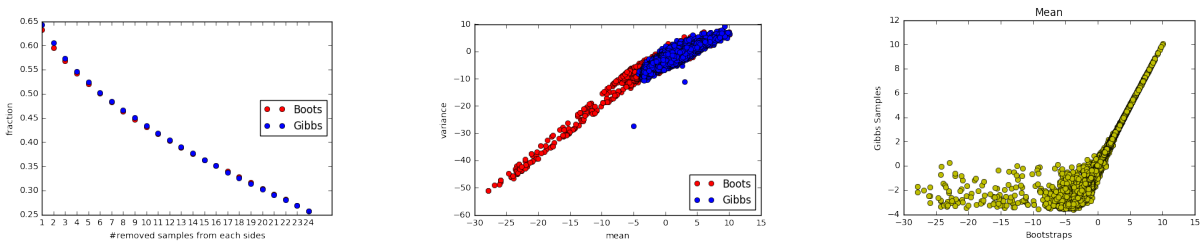


(a) Histogram of the posterior abundances of transcript 'ENST00000472311' showing Bimodality in Gibbs Sampling.



(b) Histogram of the posterior abundances of transcript 'ENST00000472311' in Bootstrap.

Figure 3: Transcript level posterior abundance plot.



(a) (b) (c)

Figure 4: Unused figures in the doc.

3 Discussion and Conclusion

We compared two resampling methods namely, bootstrap and Gibbs Sampling and We found that estimates produced by Bootstraps tend to be more conservative while that of Gibbs sample are much more liberal and assign low abundance to the transcript with fewer mapping. The reason being Bootstraps is more frequentist in its approach that is, it assigns zero probability to transcripts with no mapped reads while Gibbs sample due to the availability of prior is more bayesian in its rule and gives some probability to the transcript with no mapped read. We also observe that if a transcript with short length gets assigned with some reads can result in a big bump in overall TPM estimates which explains the difference in bootstrapping and Gibbs sampling procedure. Also, since there is a presence of uncertainty at multiple levels from sequencing to quantification it's not justifiable to use on/off switch on the transcript wrt abundance. we believe a more probabilistic view would be better which Gibbs Sampling in its current form [4] provides and can be even more optimized.

One interesting future direction is since the bootstrap in its current form does not have prior information we can make bayesian bootstrap to account for prior information. Also using uncertainty information in the form of Gibbs Sampling posterior for differential expression analysis is another awesome!!! ;p application.

Shoal: Current Progress on Empirical adaptive priors improve transcript abundance estimates in multi-sample RNA-seq data

1 Motivation:

Differential transcript expression is a challenging task in an RNA-seq experiment especially because of the presence of variance in the estimated abundances of the transcripts at multiple levels in the upstream process of quantification. Current state-of-the-art quantification tools like Salmon/kallisto does an important task of quantification which is used downstream by the various differential expression analysis tools. Errors in the differential expression analysis usually come due to type I errors (High False positive) and type II errors (High False Negative).

With the decrease in the cost of Rna-Seq experiments, multi-condition/multi-replicate data is readily available. The idea of using multi-condition data for differential expression analysis and recover from most of the type I error with a non-significant increase in the type II error has been explored deeply by many differential expression tools like Limma voom, DeSeqII. The hypothesis being if we imagine the empirical distribution of the abundances of the transcripts across replicates then to test for the differential expression we have to compare each transcripts withing replicate distribution across multiple conditions and if we can statistically differentiate (like t-test) between this two distribution then we can call a transcript differentially expressed.

In the above situation, typically type I errors would be because of the distributions being too far apart and our differential expression method is able to draw a line between them to separate them. Recent differential expression tools use multi-replicate/ condition data to correct these error by shrinking. Specifically, these tools work on the assumption that transcripts from one gene would be expressed with similar abundance across the sample and shrinks the abundances towards the mean to account for errors. Simply, in our example, we can imagine intelligently shrinking the distribution towards each other so that our differential expression the method is not able to differentiate the two distributions.

Although the above process works well in the current settings but the performance of the shrinkage after quantification would have already introduced some of the uncertainty in the analysis. In this paper we propose to shrink the estimate intelligently in the quantification process itself i.e. we use multi-condition abundance estimates as the prior to calculating per sample transcript abundance.

2 Current Plan:

- Open Questions
  - Isolator or any other tool in the comparison.
  - simulation of reads is based on EM based approach, should we simulate using VBEM?
  - learn parameter C adaptively or from multiple run of shoal.
  - Sensitivity analysis for parameter C (MITIE would be good starting point)
- Analysis

  - RSEM, kallisto, MITIE, flipflop comparison
  - Use DeSeq statistics instead of t-tests.
- Writing
  - No other tools have done the same thing (make it explicit).
  - Single Sample hypothesis correction again make it explicit.
  - TP and FN are relative, should make a comments or two.
  - typo correction on Page3
  - Define more about why use minimum phi in equivalence class relation.

3 Results:

Shoal first run Salmon on each sample separately and learn prior using transcript abundances across all samples. Later, Shoal runs Salmon's modified VBEM algorithm for quantification on each sample with the learned prior, more details can be found in Arxiv paper (have to upload). To show the difference we start with the comparison of the TPM estimates of all the transcript in one sample before and after performing inter-sample prior enhancement. As shown in fig. 5 we observe a substantial difference in the abundances specifically below 1 TPM. Also, using prior enhancement on 'VBEM' based estimation see significant improvement in the ROC curve for the simulated data.

To validate the performance we have to compare our method to currently available but the catch was all the current methods are EM-based methods, to make the comparison fair we need to run our tool on EM algorithm. On using prior directly (i.e. non-adaptive way) on shoal with Salmon EM optimization, shockingly doesn't improve the performance and we have to dig deeper and compared the estimates in one sample for real and simulated data as shown in fig. 6. Although deeper analysis on Real data is still needed but, interestingly in simulated data we observe that EM's estimate is highly sparsified which in turn was closer to truth while our flat VBEM algorithm due to it's liberal nature tends to give very small TPM's to transcript with special pattern and using prior enhancement on top makes the results sparser and much closer to EM. Which resulted in the belief that prior enhanced VBEM is sparsifying the data which is already achievable modestly by EM (w/o any prior) and makes our claim for using inter-sample information in its current form weaker.

One of the benefits of data simulation is the availability of truth but in our polyester simulation we have randomly 1-3 transcript per gene being expressed and in multiple samples, the expression can vary around a known mean for each transcript. To use the inter-sample information we need to know first that the information even available to gain at the first place? Since we know the truth we calculate ARD across each sample and their prediction based on one known truth. Once we have ARD of each transcript across every sample then we generate MARD of each transcript wrt samples and plot it against ARD of one sample. Basically, we want to test is the origin of the errors i.e. if the error being generated by either optimization algorithm (EM/VBEM) is mainly due to per sample or across sample variance, if it's across sample then only there is a hope of improvement in the estimation by using inter-sample data.

As shown in fig. 7, the color here denotes true abundances, Red is $<$1 TPM, green is 1-10 TPM and blue is $>$10TPM, x-axis is ARD in one replicate while y axis is mean of ARD across samples. As expected the Reds were clear standouts in both the optimization and there is a clear scope of improvements theoretically since there are lots of transcripts around 1.5 y-axis i.e. having big inter-replicate ARD while still marginally low one sample ARD.

Polyester simulates across replicate RNA-seq reads with some variation of expression of the transcripts. To exactly know the expression we parse the 'fastq' file of all the replicates and make a gold standard truth. The question being if we know the exact abundances of the transcript how well can our t-test differential expression method do as shown in fig. 8b. Since, the two main modes of error being FP and FN, the next question comes how much can we gain in our current estimation if we correct these errors. As the data is simulated then we can correct for FP by setting the abundance of the transcripts to 0 which are truly non-diff expressed. fig. 8a shows ROC curve with correcting every kth FP, where correcting every 2 FP takes us very close to Gold standard.

On deeper analysis, we found that there is usually three dominant modes of errors for FP. (1) Is more based on t-test threshold and usually because of very small difference in chosen threshold for diff expression. (2) 1-1-1 error: 1

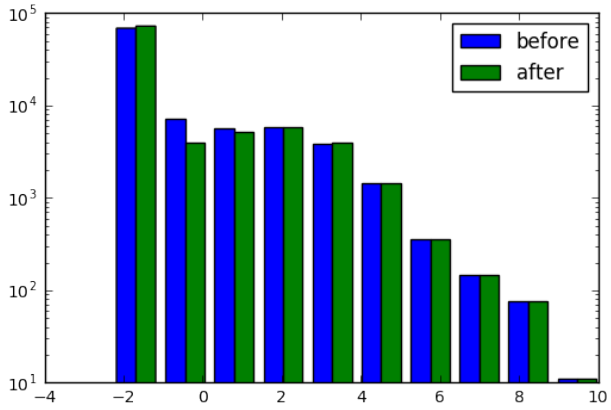
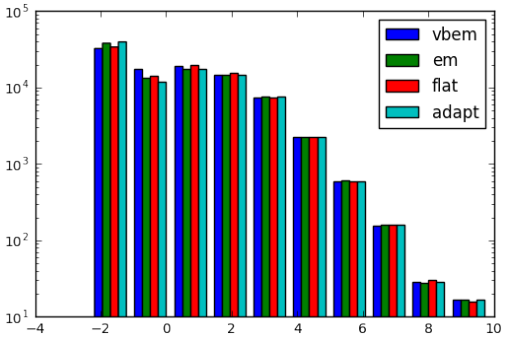
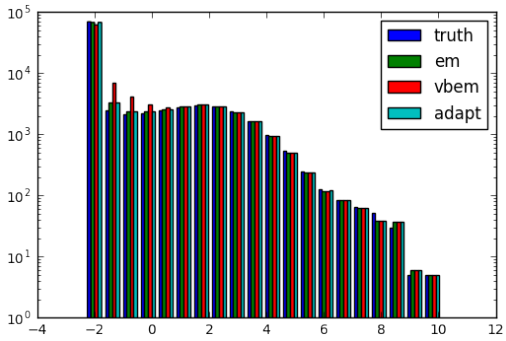


Figure 5: Histogram for the log of the abundances of the trancripts before and after using prior-enhancement on one replicate. We observe noticeable differences in the frequency of the transcript <1 TPM.

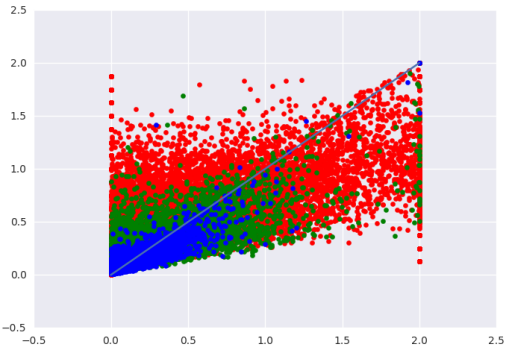


(a) Compaison of the transcript abundances in real data where vbem, em denotes no prior-enhancement and flat, adapt denotes prior-enhancement after applying flat and informative prior respectively using VBEM optimization.

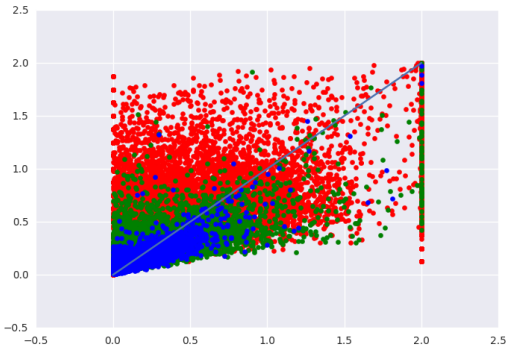


(b) Compaison of the transcript abundances in simulated data where vbem, em denotes no prior-enhancement and adapt denotes prior-enhancement after applying informative prior using VBEM optimization.

Figure 6: Comparison of abundances on various datasets led to realization that prior-enhancement on VBEM is introducing sparsity which was already achieved well by using EM w/o prior enhancement.

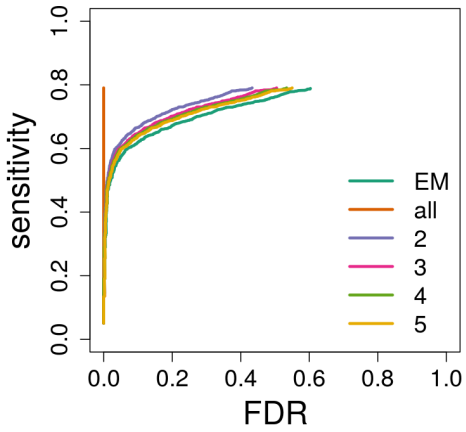


(a) VBEM: X-axis-> ARD in one replicate, Y-axis-> Mean of the ARD across samples

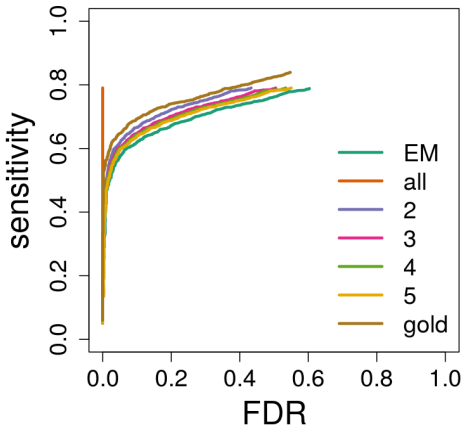


(b) EM: X-axis-> ARD in one replicate, Y-axis-> Mean of the ARD across samples

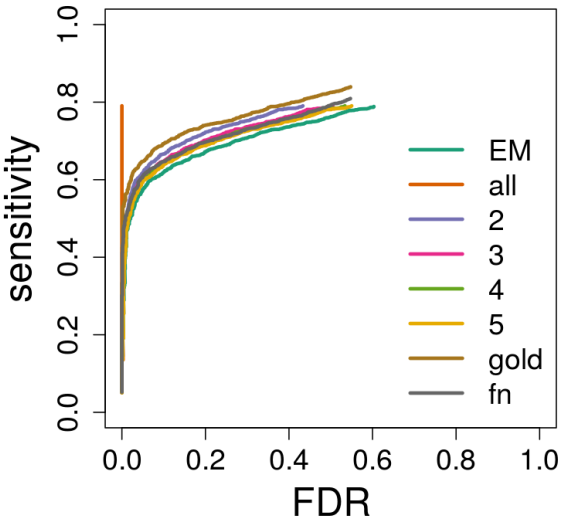
Figure 7: Comparison of the absolute relative difference of the transcript abundances.



(a)



(b)



(c)

Figure 8: ROC curve with correcting every kth False Positive. Where 'all' mean all false positive(FP) corrected and numbers means ROC curve after every kth fp transcript correction. fig. 8b adds ROC for gold standard truth i.e. true abundance generated by counting the reads in the experiment's 'fastq' file. fig. 8c shows ROC curve after we switch off the truly non expressed transcripts.

eqClass, 1 tsp and 1 read error, So equivalence classes with 1 transcript and 1 single read mapping to it within its connected component is screwing up by introducing FP since its dependant on some transcript with true diff exp. (3) This is more inferential model error which does not account for transcript with very small transcript lengths.

To test for 1-1-1 error we move on with different oracle testing. Since we know truly non-expressed transcript and if the error in diff exp is generated from giving abundance to them then we can test that by handicapping these transcript while doing EM update step and never assigning any read. But even doing this is not helpful as shown in fig. 8c.

References

[1] R. Patro, G. Duggal, M. I. Love et al. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv*, 2016. doi:10.1101/021592. URL `http://biorxiv.org/content/early/2016/08/30/021592`.

[2] W. J. Kim, J. H. Lim, J. S. Lee et al. Comprehensive analysis of transcriptome sequencing data in the lung tissues of COPD subjects. *International journal of genomics*, 2015, 2015.

[3] A. C. Frazee, A. E. Jaffe, B. Langmead et al. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 2015.

[4] E. Turro, S.-Y. Su, Â. Gonçalves et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome biology*, 12(2):1, 2011.