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.

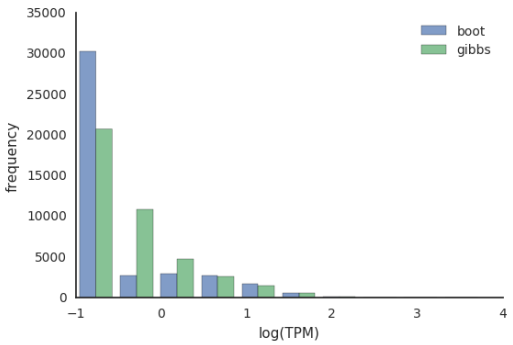
Bootstrapping and Gibbs sampling are the most prevalent resampling methods found in RNA-seq methods literature. The former is used before quantification optimization and resamples the reads from an equivalence class found by [1] while latter resamples from posterior estimates of the optimization to account 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

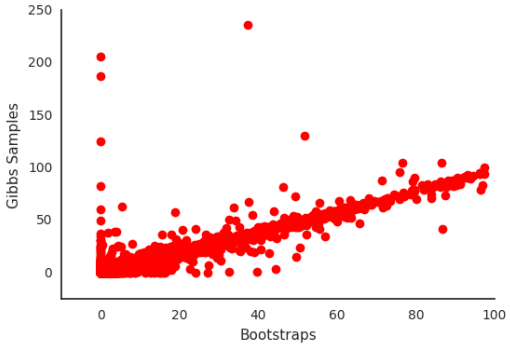
We started with one replicate and, ran bootstrap and Gibbs sampling algorithm to get 100 samples per transcript. First, the most intuitive way was to compare abundance values directly. So we took the mean abundance value of each transcript for both bootstrap and Gibbs sampling and visualize it through 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.

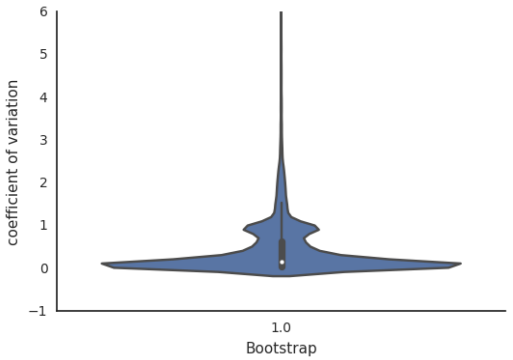


(a) Histogram of abundance values (log(TPM)) for Bootstrap and Gibbs sample in one replicate.

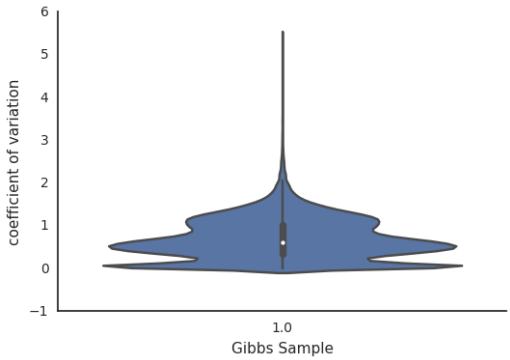


(b) Scatter plot for TPM values of Bootstrap and Gibbs Sample zoomed at 100 TPM.

Figure 1: Visualizing abundances in Real data

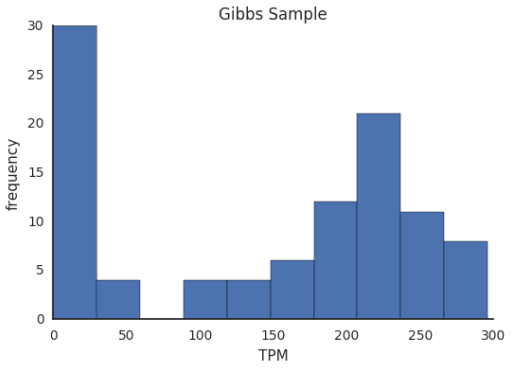


(a) Violin plot of coefficient of variation of all transcripts for Bootstrap in one one replicate.

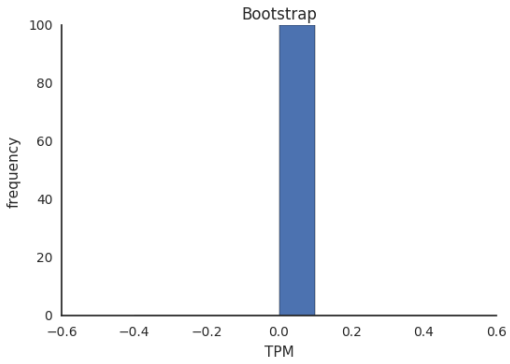


(b) Violin plot of coefficient of variation of all transcripts for Gibbs Sample in one one replicate.

Figure 2: Violin Plots on Simulated data

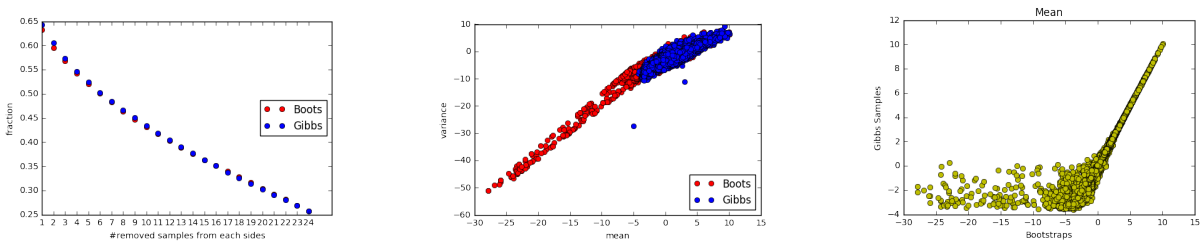


(a) Histogram of the posterior abundances of transcript 'ENST00000472311' showing Bimodality in Gibbs Sampling.



(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

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.

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.