BMC Bioinformatics

BioMed Central

Poster presentation

# Transcript quantification with RNA-Seq data

Regina Bohnert*, Jonas Behr and Gunnar Rätsch

Address: Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany

E-mail: Regina Bohnert* - Regina.Bohnert@tuebingen.mpg.de
*Corresponding author

*from* Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

*BMC Bioinformatics* 2009, **10**(Suppl 13):P5 doi: 10.1186/1471-2105-10-S13-P5

This article is available from: http://www.biomedcentral.com/1471-2105/10/S13/P5

## Motivation

Novel high-throughput sequencing technologies open exciting new approaches to transcriptome profiling. Sequencing transcript populations of interest, e.g. from different tissues or variable stress conditions, with RNA sequencing (RNA-Seq) [1] generates millions of short reads. Accurately aligned to a reference genome, they provide digital counts and thus facilitate transcript quantification. As the observed read counts only provide the summation of all expressed sequences at one locus, the inference of the underlying transcript abundances is crucial for further quantitative analyses.

## Methods

To approach this problem, we have developed a new technique, called rQuant, based on quadratic programming. Given a gene annotation and position-wise exon/intron read coverage from read alignments, we determine the abundances for each annotated transcript by minimising a suitable loss function. It penalises the deviation of the observed from the expected read coverage given the transcript weights. The observed read coverage is typically non-uniformly distributed over the transcript due to several biases in the generation of the sequencing libraries and the sequencing. This leads to distortions of the transcript abundances, if not corrected properly. We therefore extended our approach to jointly optimise transcript profiles, modeling the coverage deviations depending on the position in the transcript. Our method can be applied without knowledge of the underlying transcript abundances and equally benefits from loci with and without alternative transcripts.

## Results

To quantitatively evaluate the quality of our abundance predictions, we used a set of simulated reads from

**Table 1: Correlation of underlying expression level and inferred abundances for different approaches**

| Approach | Correlation | |
|---|---|---|
| | Across genes | Within genes (mean) |
| **Position**-wise inference with transcript profiles | 0.820 | 0.635 |
| **Segment**-wise inference with transcript profiles | 0.693 | 0.488 |
| **Position**-wise inference without transcript profiles | 0.684 | 0.540 |
| **Segment**-wise inference without transcript profiles | 0.580 | 0.367 |

rQuant, which infers transcript abundances from read data at each position, is compared against a segment-based approach, which uses averages of averaged read counts at shared transcript segments. The correlation between true and inferred abundance was determined across all annotated transcripts; for alternatively annotated genes, the average of correlation within transcripts of each gene was calculated. We compare against not optimising the transcript profiles (i.e. uniform profiles).

transcripts with known expression as a benchmark set. It was generated using the Flux Simulator [2] modeling biases in RNA-Seq as well as preparation experiments. Table 1 shows preliminary results with segment- and position-based loss as well as with and without the transcript profiles. Our results indicate that the position-based modeling together with transcript profiles allows us to accurately infer the underlying expression of single transcripts as well as of multiple isoforms of one gene locus.

## Conclusion

Our preliminary results show that modeling the transcript profiles can significantly improve the accuracy of transcript abundance estimates from RNA-Seq data. However, the described and other recent approaches [3,4] for transcript quantification with RNA-Seq rely on annotated gene structures. As most genome annotations are incomplete, they cannot reveal and quantify novel and also (novel) alternative transcripts. Nevertheless, rQuant can be extended to quantify *de novo* transcripts by combining it with a gene finding system such as mGene [5].

Revealing and quantifying novel alternative transcripts with the powerful tool of RNA-Seq will be a fundamental step towards a deeper understanding of RNA transcript regulation.

## References

1. Wang Z, Gerstein M and Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature Reviews Genetics* 2009, **10:**57–63.
2. Sammeth M: **Flux Simulator.**2009 http://flux.sammeth.net/simulator.html.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621–628.
4. Jiang H and Wong WH: **Statistical inferences for isoform expression in RNA-Seq.** *Bioinformatics* 2009, **25(8):**1026–1032.
5. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A, Krüger N, Sonnenburg S and Rätsch G: **mGene: Accurate SVM-based gene finding with an application to nematode genomes.** *Genome Research* 2009, doi:10.1101/gr.090597.108.