Behr et al. BMC Bioinformatics 2010, 11(Suppl 10):O8
http://www.biomedcentral.com/1471-2105/11/S10/O8

BMC Bioinformatics

ORAL PRESENTATION Open Access

Next generation genome annotation with mGene.ngs

Jonas Behr[1*], Regina Bohnert[1], Georg Zeller[1,2], Gabriele Schweikert[1,2,3], Lisa Hartmann[1], Gunnar Rätsch[1]

From Sixth International Society for Computational Biology (ISCB) Student Council Symposium
Boston, MA, USA. 9 July 2010

An increasingly large number of novel genomes is being sequenced and the task of automatic genome annotation has never been more important. The current revolution in sequencing technologies also allows us to obtain a detailed picture of the whole complement of expressed RNA transcripts. We have developed a novel *de novo* gene finding system mGene.ngs that combines the benefits of accurate *ab initio* gene finding with the rich information obtained in RNA sequencing (RNA-seq) experiments.

The system is based on the recently developed accurate gene finding system mGene [1], which employs state-of-the-art prediction techniques and which has been shown to perform very well compared to established gene finding systems [2]. In contrast to many HMM-based gene finders, mGene has the conceptual advantage of being very flexible in terms of incorporating heterogeneous input data. The employed inference techniques can exploit the transcriptome information already at the learning stage to appropriately adapt to the relevance of the different evidences. We show that these advantages can be translated into more accurate gene predictions. Moreover, we developed extensions of mGene.ngs to predict and quantify alternative RNA transcripts.

To provide *de novo* genome annotations based on RNA-seq experiments, we first construct a preliminary, highly specific gene set for genes that are well-covered with RNA-seq reads. In a second step, we train predictors for genomic signals on the preliminary gene set. In the third step we train mGene.ngs, using the preliminary gene models while taking advantage of the RNA-seq read coverage and genomic signal predictions.

We illustrate the power of our approach for the C. elegans genome and 50M paired-end RNA-seq reads (Illumina; 76nt). Figure 1 shows transcript level evaluation results for all annotated genes (WS200) as a function of the expression level. The *ab initio* mGene-based system (blue) trained on the annotation achieves an average transcript-level F-score of 49.9%. We achieve a slightly better performance (51.8%) for the *de novo* annotation system (green) using RNA-seq reads, but without considering the existing genome annotation. If we use the RNA-seq reads and train on the existing annotation (red), we achieve 57.6%, and can therefore take advantage of the previous annotation. We find it remarkable that for medium to high expressed genes the *de novo* gene predictions are as similar to the genome annotation as the predictions of the system, that has seen parts of the annotation in training. Comparing these results to predictions from the recently published method cufflinks [3] (black) reveals that cufflinks seems not to be able to appropriately adapt to the RNA-seq data at hand.

Investigating the contribution of individual features we found that spliced read alignments suggesting introns help most to increase the gene prediction performance; 91.6% of the achieved total improvement is due to spliced read alignments. The read coverage alone is much less informative and only leads to improvements similar to the ones achieved with transcriptome tiling arrays. We employed the developed annotation strategy for the re-annotation of the *C. briggsae* genome, for which only few transcriptome sequences are available yet. We can show that the new annotation is considerably more accurate than previous ones and additionally includes alternative RNA isoforms.

mGene.ngs will be released as open source software on http://mgene.org and is already available as Galaxy-based web-service at http://galaxy.fml.mpg.de.

* Correspondence: jonas.behr@tuebingen.mpg.de
[1]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
Full list of author information is available at the end of the article

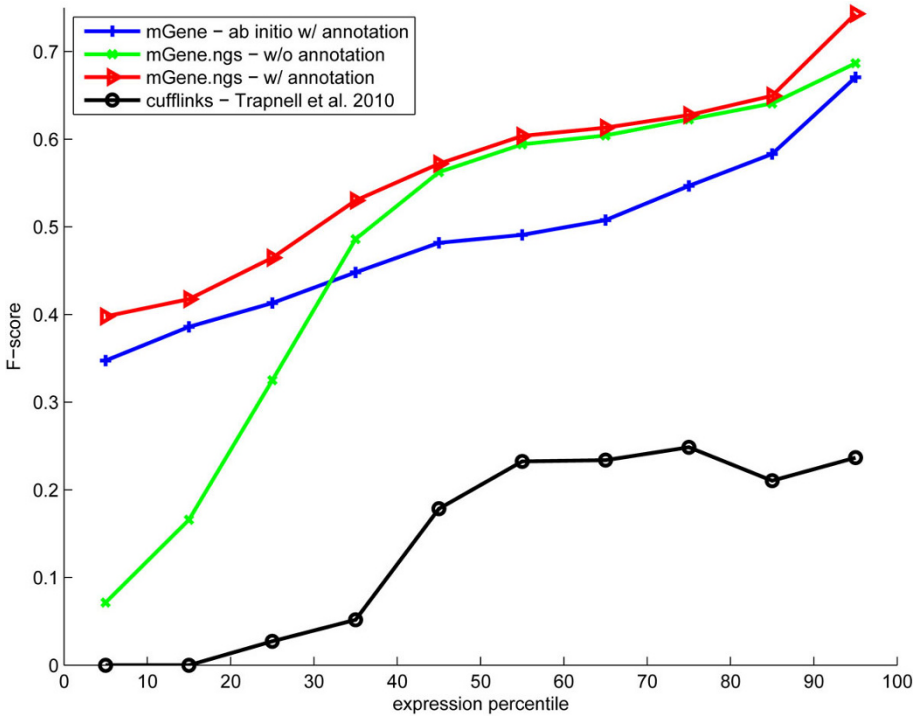

Figure 1 F-score on coding transcript level for different gene finding approaches as a function of the expression level. Coding transcript level evaluation counts a transcript as correct if all coding exons match exactly with the all coding exons of a annotated transcript. The F-score combines sensitivity and specificity penalizing large differences in these values.

Author details
[1]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. [2]Max Planck Institute for Developmental Biology, Tübingen, Germany. [3]Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Published: 7 December 2010

References

1. Schweikert , *et al*: **mGene: Accurate SVM-based gene finding.** *Genome Research* 2009, **19**:2133-2143.
2. Coghlan , *et al*: **nGASP: The nematode genome annotation assessment project.** *BMC Bioinformatics* 2008, **9**:549.
3. Trapnell , *et al*: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature Biotechnology* 2010, doi:10.1038/nbt.1621.

doi:10.1186/1471-2105-11-S10-O8
Cite this article as: Behr *et al.*: Next generation genome annotation with mGene.ngs. *BMC Bioinformatics* 2010 **11**(Suppl 10):O8.