Automated transcription start site prediction for comparative transcriptomics using the SuperGenome

Alexander Herbig[1], Cynthia Sharma[2], Kay Nieselt[1] ✉

[1]University of Tübingen, Tübingen, Germany
[2]University of Würzburg, Würzburg, Germany

Motivation and objectives

RNA deep-sequencing (RNA-Seq) has been revolutionizing eukaryotic and prokaryotic transcriptome analyses. Next-generation sequencing platforms allow one to sequence all RNA species in an organism within a couple of hours to days, and so keep accumulating massive amounts of transcriptome data. However, the bioinformatics-based analysis of this data is lagging behind. Very often, transcriptome features such as transcription start sites (TSS) and novel ncRNAs are still manually annotated, which is laborious and time-intensive. The problem is compounded for comparative transcriptomics of several species within a genus. A comparative approach would allow for refining the transcriptome annotation of the individual species by integrating the information from multiple species. This would not only lead to much better transcriptome and genome annotations but can also reveal differences in gene expression among species.

However, due to differences between the genomic architectures of the genomes, which are the result of insertions, deletions or genomic rearrangements, a direct comparison of RNA-seq data is infeasible.

Here we present two complementary approaches to solve this problem. Firstly, we developed the SuperGenome algorithm, which computes a common coordinate system for all genomes in a multiple alignment (Herbig *et al.*, 2012). The SuperGenome can be utilized for comparative analyses such as gene expression analysis, promoter sequence comparison or SNP calling. Furthermore it can be used for the comparative detection of TSS for which we - secondly - developed an automated TSS prediction method for dRNA-seq experiments (differential RNA-seq, Sharma *et al.*, 2010).

Methods

For the construction of the SuperGenome first a multiple genome alignment using *Mauve* (Darling *et al.*, 2004) is computed. Based on this alignment, the SuperGenome as a common genomic coordinate system and a mapping of each position of each single genome to a position in the SuperGenome is calculated. Next, all genome-specific data can be mapped to the common coordinate system, which includes genomic annotations or sequence information but also expression values derived from mapped read data, thus making a direct comparison of these data possible.

Our automated TSS prediction approach consists of several steps: first an initial detection of TSS in each genome (species) used in the experiment is conducted. Here, positions are localized, where a significant number of reads start (in comparison to local background). To evaluate if the reads starting at this position originate from primary transcripts, the enrichment factor is calculated by comparing the data from the standard library with a library that has been treated with terminator exonucleases (TEX), which specifically degrades processed RNAs with a 5′-mono-phosphate (Sharma *et al.*, 2010). For all positions where these values exceed the thresholds a TSS candidate is called. In the next step the TSS candidates of each species are mapped to the SuperGenome to assign each TSS to the corresponding TSS in the other species. Finally, all TSS are then characterized on the SuperGenome level with respect to their occurrence in the different species.

Results and discussion

Here we present the application of our SuperGenome approach and TSS prediction method to an RNA-seq experiment using four *Campylobacter jejuni* strains.

Mapping of RNA-seq data into the SuperGenome allows for a direct comparison of expression patterns among the four strains, e.g. for a visual comparison in a genome browser (Figure 1). In combination with our TSS prediction

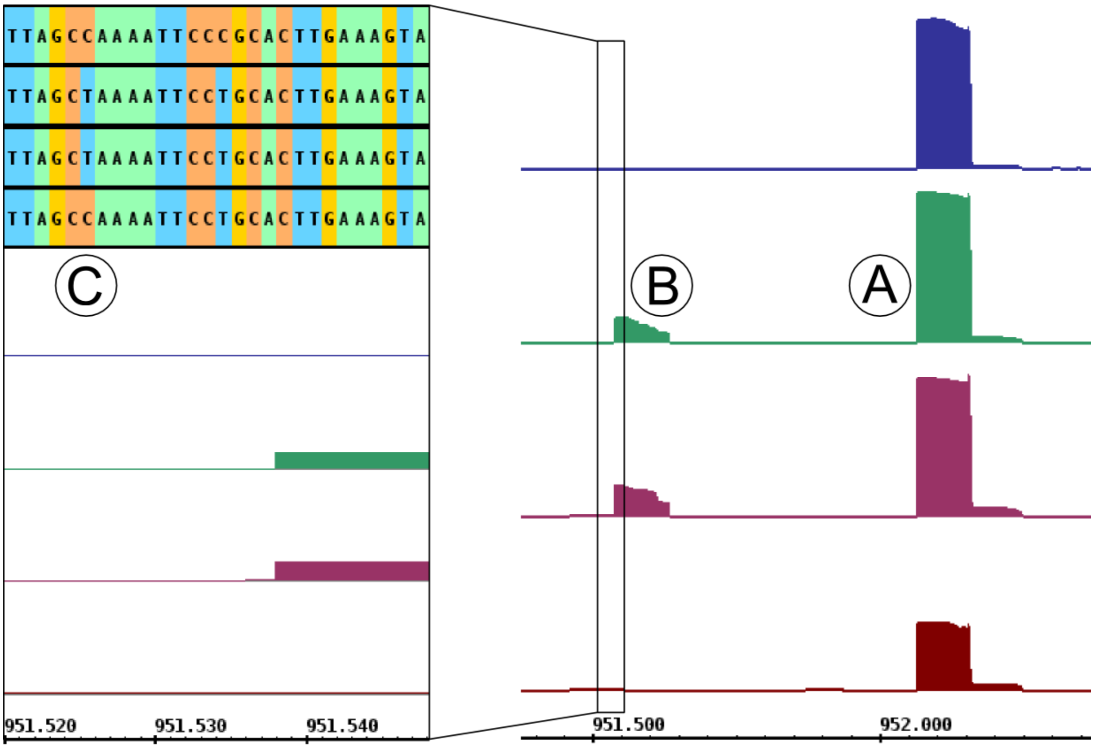

Figure1. Visualization of RNA-seq expression data of four C. jejuni strains. The graphs have been projected into the SuperGenome coordinate system (bottom). A: Expression pattern indicating a TSS that is conserved in all strains. B: Expression pattern showing a TSS that is only conserved in two of the strains. A close-up view of the respective promoter region is shown on the left. C: Alignment information from the SuperGenome reveals a SNP in the promoter region, which possibly induces the strain-specific expression pattern.

method, we are thus able to identify TSS in several species simultaneously and classify on the SuperGenome level whether they are detected in all strains (Figure 1A) or whether they are specific for only a subset of strains (Figure 1B).

Genome-wide application of our automated TSS prediction resulted in the annotation of more than 3000 TSS of which more than 1000 were detected in all four strains.

In addition, the SuperGenome allows for a comparative analysis of promoter regions related to the detected TSS. By this means, variations in the promoter sequence can easily be identified potentially explaining differences in the transcriptomic architectures of the investigated organisms (Figure 1C). Combining these information can help to elucidate novel mechanisms of transcriptional regulation and explain phenotypic diversity, e.g., in the context of pathogenicity. Overall our high-resolution transcriptome map revealed regulatory elements and their conservation in multiple C. *jejuni* strains on a genome-wide scale.

In summary, our TSS prediction procedure in combination with the SuperGenome provides a novel approach to comparative analysis of RNA-seq data, facilitating the cross-genome annotation of transcriptome features such as TSS maps and promoter regions.

References

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**(7): 1394-403. doi:10.1101/gr.2289704

Herbig A, Jäger G, Battke F, Nieselt K (2012) GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics.* **28**(12): i7-15. doi:10.1093/bioinformatics/bts217

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature.* **464**(7286): 250-5. doi:10.1038/nature08756