Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genome analysis

Necklace: combining reference and assembled transcriptomes for RNA-Seq analysis

Nadia M Davidson[1,2,*] and Alicia Oshlack[1,2,*]

[1]Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia, [2]School of Bio-Sciences, University of Melbourne, Victoria, Australia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: RNA-Seq analyses can benefit from performing a genome-guided and de novo assembly, in particular for species where the reference genome is incomplete. However, tools to integrate assembled transcriptome with reference annotation are lacking.
Results: Necklace is a software pipeline to run genome-guided and de novo assembly and combine the resulting transcriptomes with reference genome annotations. Necklace constructs a compact but comprehensive superTranscriptome out of the assembled and reference data. Reads are subsequently aligned and counted in preparation for differential expression testing.
Availability: Necklace is available from https://github.com/Oshlack/necklace/wiki under GPL 3.0.
Contact: nadia.davidson@mcri.edu.au or alicia.oshlack@mcri.edu.au
Supplementary information: Supplementary methods are available at *Bioinformatics* online.

1 Introduction

Despite the increasing number of species with a sequenced genome, the vast majority of reference genomes are incomplete. They may contain gaps, have unplaced assembly scaffolds and be poorly annotated. Therefore, analyzing RNA-Seq data using just the reference genome has the potential to miss important biology for many organisms. Ideally, an RNA-Seq analysis could utilise prior knowledge of the gene-models available from a reference genome and annotation, whilst also extracting information about the genes from the data itself through genome-guided and/or de novo assembly (Martin and Wang, 2011).

In (Davidson *et al.*, 2017) we introduced the concept of the superTranscriptome, where each gene is represented by one sequence containing all of that gene's exons in transcriptional order. SuperTranscripts provided a convenient means in which transcriptomes from difference sources, such as assembly and annotation, can be combined into a compact and unified reference. When applied to chicken, we showed that we could recover hundreds of segments of genes that were absent from the chicken reference genome.

Here we present software called necklace which automates the process described in (Davidson *et al.*, 2017) for any species with an incomplete reference genome. Necklace takes as input a configuration file containing paths to the RNA-Seq reads, a reference genome and one or more reference genome annotation. Because de novo assembly is error prone, we require that any gene discovered through de novo assembly be found amongst the coding sequence of a related (well annotated) species. Therefore, the genome and annotation of a related species must also be provided to necklace. Necklace will then run the steps involved in genome-guided and de novo assembly, and combine the assembled transcriptome with reference annotations for the species of interest. After building the superTranscriptome, necklace will align and count reads in preparation for testing for differential gene expression and differential transcript usage analysis.

We applied necklace to RNA-seq analysis of sheep milk and found 18% more reads could be assigned to genes and 19% more differentially expressed genes were detected using the necklace derived superTranascriptome reference compared with the reference genome on its own.

2 Methods

Necklace is a pipeline constructed using the bpipe framework (Sadedin *et al.*, 2012). It steers external software, such as aligners and assemblers, as well as a set of its own utilities, written in c/c++. Necklace consists of several sequential stages: initial assembly, clustering transcripts into gene groupings, reassembly to build the superTranscriptome and finally alignment and counting of mapped reads in preparation for differential expression testing. Each of these sequential stages consists of several sub-stages (Figure 1A):

1. **Assembly** – The assembly stage creates three different transcriptomes. First reads are aligned to the reference genome using HISAT2 (Kim *et al.*, 2015) and genome-guided assembly is performed with StringTie (Pertea *et al.*, 2015). This assembly is combined with reference annotations and then flattened based on genomic location, so that each exon is reported only once and overlapping exons are merged. Exonic sequence is then extracted from the genome and concatenated to build a “genome-based” superTranscriptome. In parallel, the related specie’s annotation is used to create a “genome-based” superTranscriptome (without genome-guided assembly). Finally, RNA-Seq reads are de novo assembled with Trinity (Haas *et al.*, 2013).
2. **Clustering** – This step assigns de novo assembled transcripts to gene clusters prior to building the final superTranscriptome. Those assemblies aligning to the genome-based superTrascriptome (using Blat (Kent, 2002)) are allocated to known genes while those not aligning to the genome, but found in the related species superTranscriptome are assigned to novel genes. De novo assembled transcripts that align to more than one gene are removed to avoid false chimeras (Davidson and Oshlack, 2014) from being introduced into the superTranscriptome.
3. **Reassembly** - Each cluster consists of a gene’s genome-based superTranscript and/or its set of de novo assembled transcripts. The transcripts in each cluster are merged together through Lace assembly (Davidson *et al.*, 2017), to produce one superTranscript per gene.
4. **Summarization** - Reads are aligned back to the superTranscriptome using HISAT2 and fragments counted per gene using featureCounts (Liao *et al.*, 2014). Splice junctions reported by HISAT2 are used to segment each superTranscript into a set of contiguous “blocks”. Fragments are then counted in “blocks” and can be used for differential isoform detection like exon counts.

3 Application

To demonstrate the utility of necklace, we applied it to public RNA-Seq from Churra sheep milk and compared transcriptome expression at day 10 to day 150 post lambing (Suárez-Vega *et al.*, 2016). Necklace was given the sheep reference genome, Oar_v3.1. Human, with the hg38 reference genome, was used as the related species. For both genomes, version 90 of the Ensembl annotation was used. Compared to the Ensembl sheep annotation, necklace annotated 76% more bases and 8% more genes. This more comprehensive reference resulted in 18% more reads being assigned to genes by featureCounts compared to the reference annotation alone, which in turn lead to the identification of more significantly differentially expressed genes (456 compared to 383, FDR<0.05). Some of this difference could be attributed to the inclusion of novel unannotated genes (66), but necklace also improved the differential expression detection amongst known genes (for example, *SERTM1,* Figure 1B).

Funding

AO is funded by an *NHMRC CDF GNT1126157.*

Conflict of Interest: none declared.

Acknowledgements: would like to thank Anthony Hawkins, the author of Lace.

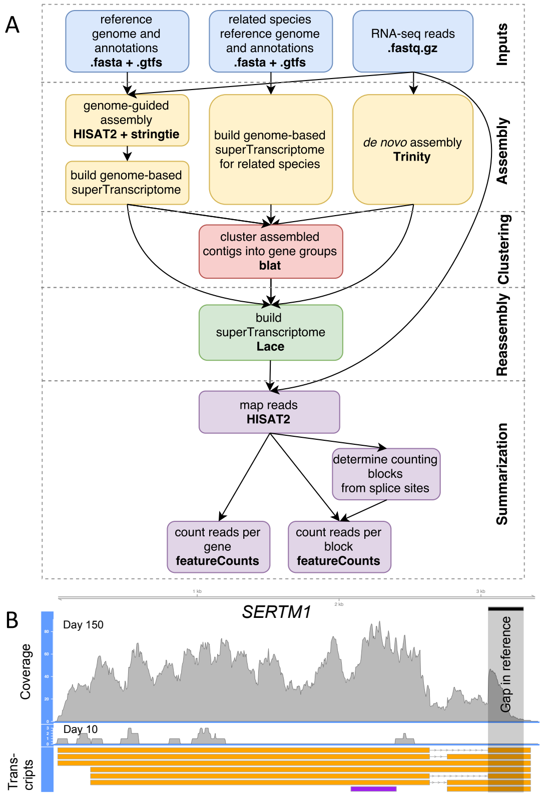

Fig. 1. A) An overview of the necklace pipeline. External software that necklace runs is shown in bold. B) Read coverage aggregated over samples for the necklace assembled superTranscript of *SERTM1*. This gene is found to be significantly differentially expressed using the necklace generated reference, but is missed when the reference genome and annotation are used in isolation due to low read counts. The reference annotation consists of a single transcript of 321bp (shown in purple), whereas the de novo assembled gene consists of seven transcripts up to 3333bp long (shown in orange) and includes approximately 250bp that is absent from the reference genome, in a location consistent with a genome assembly gap. The genome-guided transcripts that were assembled for this gene were filtered out by stringTie’s merge function due to an average FPKM < 1.

References

Davidson,N.M. *et al.* (2017) SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol. 2017 181*, **18**, 148.

Haas,B.J. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–512.

Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, **12**, 656–64.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–60.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–30.

Martin,J. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290.

Sadedin,S.P. *et al.* (2012) Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, **28**, 1525–6.

Suárez-Vega,A. *et al.* (2016) Comprehensive RNA-Seq profiling to evaluate lactating sheep mammary gland transcriptome. *Sci. Data*, **3**, 160051.