MOLECULAR ECOLOGY
RESOURCES

Reference-free transcriptome assembly in non-model animals from next generation sequencing data

Journal:	*Molecular Ecology Resources*
Manuscript ID:	MER-12-0016.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	06-Mar-2012
Complete List of Authors:	Cahais, Vincent; Université Montpellier 2, CNRS UMR 5554 Gayral, Philippe; CNRS, Institut des Sciences de l'Evolution; Université Montpellier 2, CNRS UMR 5554 Tsagkogeorga, Georgia; Université Montpellier 2, CNRS UMR 5554 Melo-Ferreira, José; CIBIO, Centro de Investiga鈦o em Biodiversidade e Recursos Geneticos, Ballenghien, Marion; Université Montpellier 2, CNRS UMR 5554 Weinert, Lucy; Université Montpellier 2, CNRS UMR 5554 Chiari, Ylenia; University of Montpellier 2, Institut des Sciences de l'Évolution; Belkhir, Khalid; Université Montpellier 2, CNRS UMR 5554 Ranwez, Vincent; Université Montpellier 2, CNRS UMR 5554 Galtier, Nicolas; Université Montpellier 2, CNRS UMR 5554
Keywords:	Bioinfomatics/Phyloinfomatics, Transcriptomics, Population Genetics - Empirical, Invertebrates

Reference-free transcriptome assembly in non-model animals from next generation sequencing data

Cahais V.[1], Gayral P.[1], Tsagkogeorga G.[1,2], Melo-Ferreira J.[3], Ballenghien M.[1], Weinert L.[1,4], Chiari Y.[1,3], Belkhir K.[1], Ranwez V.[1], Galtier N[1*].

[1] Université Montpellier 2, CNRS UMR 5554, Institut des Sciences de l'Evolution de Montpellier, Place E. Bataillon, 34095 Montpellier, France;

[2] School of Biological & Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS

[3] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

[4] Medical Research Council (MRC), Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK.

Corresponding author:
Nicolas Galtier
CNRS UMR5554 – Institut des Sciences de l'Evolution
Place E. Bataillon – CC64 – 34095 Montpellier, France
Phone: (+33) 467 14 48 18 Fax: (+33) 467 14 36 10
nicolas.galtier@univ-montp2.fr

Key words: NGS, transcriptomics, paralogues, alleles
Running title: *de novo* NGS-based transcriptome assembly

Abstract

Next-generation sequencing technologies offer the opportunity for population genomic study of non-model organisms sampled in the wild. The transcriptome is a convenient and popular target for such purposes. However, designing genetic markers from NGS transcriptome data requires assembling gene coding sequences out of short reads. This is a complex task owing to gene duplications, genetic polymorphism, alternative splicing, and transcription noise. Typical assembling programs return thousands of predicted contigs, whose connection to the species true gene content is unclear, and from which SNP definition is uneasy. Here the transcriptomes of five diverse non-model animal species (hare, turtle, ant, oyster, tunicate) were assembled from newly-generated 454 and Illumina sequence reads. In two species for which a reference genome is available, a new procedure was introduced to annotate each predicted contig as either a full length cDNA, fragment, chimera, allele, paralogue, genomic sequence or other, based on the number of, and overlap between, BLAST hits to the appropriate reference. Analyses showed that (i) the highest-quality assemblies are obtained when 454 and Illumina data are combined, (ii) typical *de novo* assemblies include a majority of irrelevant cDNA predictions, and (iii) assemblies can be appropriately cleaned by filtering contigs based on length and coverage. We conclude that robust, reference-free assembly of thousands of genes from transcriptomic next-generation sequence data is possible, opening promising perspectives for transcriptome-based population genomics in animals. A Galaxy pipeline implementing our best-performing assembling strategy is provided.