IVYSPRING INTERNATIONAL PUBLISHER

Journal of Genomics

2017; 5: 54-57. doi: 10.7150/jgen.19504

Research Paper

De Novo Assembly of the Liver Transcriptome of the European Starling, *Sturnus vulgaris*

Mark F. Richardson[1, 3]✉, William B. Sherwin[2, 4], Lee A. Rollins[2, 3]

1. Deakin University, Bioinformatics Core Research Group, 75 Pigdons Road, Locked Bag 20000, Geelong, VIC 3220, Australia;
2. Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia;
3. Deakin University, School of Life and Environmental Sciences, Centre for Integrative Ecology, 75 Pigdons Road, Locked Bag 20000, Geelong, VIC 3220, Australia;
4. Cetacean Research Unit, Murdoch University, South Road, Murdoch, Western Australia 6150, Australia.

✉ Corresponding author: Mark F. Richardson, Email: m.richardson@deakin.edu.au

Published: 2017.05.04

Abstract

The European starling, *Sturnus vulgaris*, is a prolific and worldwide invasive species that also has served as an important model for avian ecological and invasion research. Although the genome sequence recently has become available, no transcriptome data have been published for this species. Here, we have sequenced and assembled the *S. vulgaris* liver transcriptome, which will provide a foundational resource for further annotation and validation of the draft genome. Moreover, it will be important for ecological and evolutionary studies investigating the genetic factors underlying rapid evolution and invasion success in this global invader.

Key words: *Sturnus vulgaris*, European starling, *de novo* transcriptome assembly, invasive species, RNA-seq.

Introduction

The European starling is one of the best-studied avian species, yet no genomic resources existed prior to the recent draft genome (December 2015) and no transcriptome data sets are currently available. Starlings are also an important species in invasion genetics research, due to their global introduction history [1], their invasion success and their known patterns of morphological and genetic variation across introduced ranges [2–4], making them ideal for studies of rapid evolution replicated across invasions. However, these studies require improved genetic resources.

Here we characterise the first starling transcriptome data set, produced using liver tissue from individuals sourced from the range edge in Western Australia and carrying genetic variants previously shown to be under selection [4]. Two juvenile male starlings (S274, S290) were collected from Western Australia (S274: S 33.81542, E 120.95964; S290: S 33.79625, E 120.90078). Total RNA was extracted and mRNA libraries prepared following [5]. Libraries were barcoded and run together on one lane of HiSeq 2500 (Illumina Inc, San Diego, USA), generating ~230 million 2 x 125 bp paired-end reads.

Raw reads were pooled and processed using Trimmomatic v0.33 [6] using the following parameters, ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:4 HEADCROP:13 AVGQUAL:30 MINLEN:36, reducing the dataset to ~36 million paired reads and ~9 million 'orphaned' reads (Table 1). We used both the filtered paired-end and 'orphan' reads in the subsequent *de novo* assembly with Trinity v2.1.1 [7] (using default settings and retaining only those contigs > 300bp) producing 59,557 transcripts, encompassing 48,279 unigenes. After assembly, filtered paired-end reads were mapped to the assembled transcripts with

Bowtie2 v2.2.4 [8] (82% of reads map back to the *de novo* assembly; 69 % are properly paired), gene expression was estimated using RSEM v1 2.14 [9] (mean transcript expression of 10.7 FPKM) and open reading frames (ORFs) were predicted for all assembled transcripts using Transdecoder (https://github.com/TransDecoder/TransDecoder).

Table 1. Transcriptome assembly and annotation statistics compared to other passerine transcriptomes.

	S. vulgaris[a]	*C. chloris*[b]	*Z. albicollis*[c]
Raw sequencing reads	230403632	nr	-
Reads used in assembly	45309889	~500000000	-
Unfiltered *de novo* assembly			
Number of unigenes	48279	nr	-
Number of transcripts	59557	66072	313060
n50 transcript length (bp)	1765	803	3979
sum transcript length (Mb)	64993660	39395826	334636954
median transcript length (bp)	626	367	345
mean transcript length (bp)	1091	596	1069
GC %	48.28	46.34	45.43
Filtered *de novo* assembly			
Number of unigenes	18678	-	-
n50 longest unigene/transcript	2232	-	-
Sum longest unigene/transcript	26825376	-	-
Median longest unigene/transcript length (bp)	979	-	-
Mean longest unigene/transcript length (bp)	1436	-	-
Number of transcripts	23945	-	-
n50 transcript length (bp)	2328	-	-
Sum transcript length (Mb)	37637538	-	-
Median transcript length (bp)	1178	-	-
Mean transcript length (bp)	1572	-	-
Annotation statistics			
Unfiltered transcriptome			
Transcripts with Blastx match	33041 (55%)	nr	-
Transcripts with Blastp match	24715 (41%)	23,151 (35%)	-
Transcripts with GO terms	27576 (46%)	nr	-
Filtered transcriptome			
Transcripts with Blastx match	19701 (82%)	-	-
Transcripts with Blastp match	17898 (75%)	-	-
Transcripts with GO terms	17462 (73%)	-	-

[a]This study; [b]data from Meitern *et al.* (2014); [c]data calculated from NCBI GBBC00000000.1; nr, not reported; percentage in parentheses

Transvestigator v0.1 [10] was used to filter and prepare the raw transcriptome assembly for NCBI Transcriptome Shotgun Assembly (TSA). Briefly, we filtered poorly supported and lowly expressed transcripts if they had a transcript per million (TPM) value < 0.5 (7,873 transcripts) and an isoform expression level less than 5% of the parent unigenes expression (7,708 transcripts). Additionally, only those transcripts passing the above filter and containing a predicted ORF were kept (18,103 transcripts removed). Assembled sequences were queried against the UniVec database (Accessed January 2017) and any sequences with a strong match were removed (1/1,000,000 chance of a random match for queries of 350 Kb, terminal match score ≥ 24, internal match score ≥ 30). Lastly, preparation for TSA submission included ensuring there was only one ORF per transcript, the ORF was on the positive strand and contained within transcript coordinates, and that start and stop codons were properly created. The final filtered transcriptome contains 23,945 protein-coding transcripts from 18,678 'unigenes' (accessible via the NCBI TSA accession: GFDQ00000000). Overall, the filtering resulted in the removal of 1,545 transcripts that may have contained some protein-coding potential. As non-coding RNAs have key biological roles (rRNA, tRNA etc), may be significant in terms of adaptive processes and are useful for genome annotation we also provide the unfiltered transcripts (includes those not passing ORF filters, but meeting support and expression thresholds and those with ORFs not meeting expression thresholds) in the associated data repository [13]. Functional annotation was conducted utilising the Trinotate pipeline (http://trinotate.sourceforge.net) – these are also provided in the associated repository.

To assess the relative quality and completeness of the assembly, we compared core assembly statistics and evaluated completeness using the BUSCO (benchmarking universal single-copy orthologs) vertebrate gene set [11] to two recent passerine transcriptomes available in the NCBI TSA that also used the Trinity assembler: *Carduelis chloris*, GBCG00000000.1 [12] and *Zonotrichia albicollis*, GBBC00000000.1 (Table 1, 2). Our assembly compares favorably in terms of core statistics (Table 1). Similarly, BUSCO assessment also reveals comparable completeness (*S. vulgaris*, 50% complete; *C. chloris*, 30% complete; *Z. albicollis*, 62% complete; Table 2) even with a much smaller number of transcripts assembled.

The recent *S. vulgaris* draft genome assembly (NCBI Bioproject: PRJNA304638) includes a predicted transcriptome produced by the NCBI Eukaryotic Genome Annotation Pipeline. We compared the final

de novo assembled liver transcriptome set (valid ORF and evidence of expression) to the draft genome and predicted transcriptome to aid characterization of gene features and validate gene models through both a standard and reciprocal best-hit BLAST approach and mapping of our mRNA reads to the genome. As our transcriptome assembly is based on a single tissue we only expect a partial representation of the predicted genes to be confirmed. The standard BLASTn approached revealed 23,557 (98.4 %) of our final transcripts had a significant match (e-value $<10^{-5}$) to the genome, with 20,768 (86.7 %) having a significant match to the predicted transcriptome. The reciprocal best-hit approach reduced the number of significant matches to the predicted transcriptome to 10,265 (42.9%), which may in part be due to predicted protein coding genes mapping to multiple isoforms present in our transcriptome assembly. While 11.7% (2,790) of our final transcripts did not produce a significant match to the predicted transcriptome they did produce a significant match to the draft genome assembly – underlying the importance of including RNA-Seq data in the annotation process. We used bbmap v35, (https://sourceforge.net/projects/bbmap/) with default settings to map the mRNA reads used to construct the *de novo* assembly to both the draft genome and predicted transcriptome, with 84.2% and 69.1% of reads mapping respectively. The higher percentage of reads mapping to the genome indicates there may be missed features in the predicted transcriptome. Additionally, standard BLASTn of the unfiltered *de novo* assembly to the genome revealed that 58,192 (97.8%) transcripts had a significant match, whereas only 38,021 (63.8%) had a significant match to the predicted transcriptome. This gives further support to the previous conclusion of missing features in the predicted transcriptome and suggests that our filtering (only keeping transcripts with a valid ORF and evidence of expression) may have removed some true transcripts (potential ncRNAs) that are encoded in both the genome and predicted transcriptome. Consequently, we have included these transcripts in the associated public repository [13].

Table 2. BUSCO evaluations of completeness against the vertebrate gene set compared to other passerine transcriptomes.

	S. vulgaris	*C. chloris*	*Z. albicollis*
Complete	1523 (50%)	904 (30%)	1860 (62%)
Single	1409 (47%)	900 (30%)	1563 (52%)
Multi	114 (4%)	4 (~0%)	297 (10%)
Fragment	346 (11%)	361 (12%)	249 (8%)
Missing	1154 (38%)	1758 (58%)	914 (30%)

This resource will be useful for the validation of gene models predicted in the recent draft genome assembly and for future research into the physiology and ecology of this and other closely related species. In particular, these data will enable a greater understanding of the genetic factors underlying rapid evolution and invasion success in this global invader.

Availability of supporting data

The datasets supporting the results presented here are available at http://dro.deakin.edu.au/view/DU:30091025 and doi: 10.4225/16/5893999965ac2 [13]. All raw sequencing data used in this study is available in the NCBI SRA and associated with the BioProject accession, PRJNA335913. The final transcriptome assembly has been deposited at DDBJ/EMBL/GenBank under the accession GFDQ00000000. The version described in this paper is the first version, GFDQ01000000.

Acknowledgements

We thank Peri Bolton for use of the starling image in our graphical abstract. This work was supported by funding from the Centre for Integrative Ecology at Deakin University, an ARC DECRA Fellowship to LAR and an ARC Linkage Project (0455776) to WBS. We thank Dan Selechnik for conducting RNA extractions for this study.

Competing Interests

The authors have declared that no competing interest exists.

References

1. Rollins LA, Woolnough AP, Sherwin WB: Population genetic tools for pest management: a review. Wildl Res 2006, 33:251–261.
2. Berthouly-Salazar C, Hui C, Blackburn TM, Gaboriaud C, van Rensburg BJ, van Vuuren BJ, Le Roux JJ: Long-distance dispersal maximizes evolutionary potential during rapid geographic range expansion. Mol Ecol 2013, 22:5793–804.
3. Cardilini APA, Buchanan KL, Sherman CDH, Cassey P, Symonds MRE: Tests of ecogeographical relationships in a non-native species: what rules avian morphology? Oecologia 2016, 181:783–793.
4. Rollins LA, Woolnough AP, Fanson BG, Cummins ML, Crowley TM, Wilton AN, Sinclair R, Butler A, Sherwin WB: Selection on Mitochondrial Variants Occurs between and within Individuals in an Expanding Invasion. Mol Biol Evol 2016, 33:995–1007.
5. Rollins LA, Richardson MF, Shine R: A genetic perspective on rapid evolution in cane toads (Rhinella marina). Mol Ecol 2015, 24:2264-2276.
6. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014, 30:2114–20.
7. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011, 29:644–52.
8. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9:357–9.
9. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011, 12:323.
10. DeRego T, Hall B, Tate R, Geib S: Transvestigator early release. ZENODO 2014, doi:10.5281/zenodo.10471

11. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015, 31:3210–2.
12. Meitern R, Andreson R, Hõrak P: Profile of whole blood gene expression following immune stimulation in a wild passerine. BMC Genomics 2014, 15:533.
13. Richardson MF, Sherwin WB, Rollins LA: Supporting data for "*De novo* assembly of the liver transcriptome of the European starling, *Sturnus vulgaris*". doi: 10.4225/16/5893999965ac2.