Title

De novo transcriptome assembly of the mountain fly *Drosophila nigrosparsa* using short RNA-seq reads

Authors

Wolfgang Arthofer[a*†], Francesco Cicconardi [a*†], Nicola Palmieri[b†], Viola Nolte[b], Christian Schlötterer[b], Birgit C. Schlick-Steiner[a‡], Florian M. Steiner [a‡]

Affiliations

[a] Molecular Ecology Group, Institute of Ecology, University of Innsbruck, Technikerstraße 25, 6020 Innsbruck, Austria

[b] Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria

[*] Corresponding author (email: wolfgang.arthofer@uibk.ac.at)

[†] Equally contributing first authors

[‡] Equally contributing senior authors

Running title

Drosophila nigrosparsa transcriptome

Keywords

Drosophila nigrosparsa, transcriptome, *de novo* assembly, Illumina

Introduction

With 1,178 species, distributed over the entire globe (http://www.taxodros.uzh.ch/lists/SPECIES-LIST_GE_SG; retrieved 25 June 2014), the genus *Drosophila* is an excellent system for cross-taxon comparisons. The genus comprises widespread generalist and highly specialised species differing in ecological niches and habitat requirements, both of which are well described for several species (e.g., Bächli & Burla 1985). *Drosophila* is considered the best characterised model in animal genetics (Ashburner *et al*. 2005, http://flybase.org/) and allows to link genetic data with essential phenotypic information from a vast range of disciplines, including developmental biology, cell biology and physiology. The GOLD database (http://genomesonline.org, retrieved 25 June 2014) currently lists 21 whole genome projects for the genus *Drosophila* as "complete or draft" and additional 16 projects as "in progress". For the model species *D. melanogaster*, a vast set of transcriptome data from different developmental stages has been published recently (Brown *et al.* 2014).

Drosophila (*Drosophila*) *nigrosparsa* is a habitat specialist restricted to the European montane/alpine zone (Bächli 2008). Mountain biodiversity is considered highly vulnerable to ongoing climate warming (IPCC 2013), and organisms at high altitudes have only limited possibility to shift to cooler habitats at elevations above (Pertoldi & Bach 2007). For such species, rapid evolution may offer a solution for long-term survival. We are establishing *D. nigrosparsa* as a model system to test the extent and tempo of adaptive evolution under thermal stress in the laboratory.

In this study, we used Illumina high-throughput sequencing to assemble the species' transcriptome using the pooled mRNA from 22 developmental and physiological stages (eggs: untreated; 1st instar larvae: untreated; 2nd instar larvae: untreated, heat shocked, heat shocked after heat hardening, cold shocked, cold shocked after cold hardening; 3rd instar larvae: untreated; pupae: untreated; freshly eclosed females: untreated; freshly eclosed males: untreated; adult females: same treatments as 2nd instar larvae; adult males: same treatments as 2nd instar larvae). For extraction of total RNA using TriFast Reagent (peqlab, Erlangen, Germany), we used two pooled individuals of each stage, except for eggs and 1st instar larvae, where each 150 individuals were pooled. Equal amounts of RNA from each stage were combined for the construction of a single, pooled library for sequencing. Our aim was the reconstruction of the transcriptome with a discrete functional coverage and complexity. Using 140 million reads, we were able to report a high number of genes with different isoforms per locus and to achieve long transcripts compared with the other available *Drosophila* assemblies. Our study also confirms *D. nigrosparsa*'s taxonomic position within the *Drosophila* subgenus and the high similarity to *D. grimshawi* and *D. virilis*.

The data presented here are the first genomic resource available for the mountain fly *D. nigrosparsa* and will facilitate future research on adaptive evolution. They are also valuable for transcriptome-based phylogenetic reconstruction of the subgenus *Drosophila*.

Data Access

- *NGS sequence data:* NCBI BioProject PRJNA232118
- *Trinity transcripts:* Dryad doi:10.5061/dryad.0mv56
- *Annotated transcripts:* Dryad entry doi:10.5061/dryad.0mv56

Meta Information

- *Sequencing center:* BGI Hong Kong Co., Ltd (http://www.genomics.cn/en/index)
- *Platform and model:* Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA)

- *Design description:* The goal of our study was to assemble a highly accurate transcriptome of *D. nigrosparsa*, an alpine/montane species that we are currently developing towards a model system to test the extent and tempo of adaptive evolution under thermal stress. We sampled individuals at different developmental and physiological stages with special focus on varying exposure temperatures, and pooled equal amounts of RNA for the construction of a non-normalized library.
- *Analysis type:* cDNA
- *Run date:* 27 December 2011

Library

- *Strategy:* 100 bp paired-end Illumina sequencing of non-normalized cDNA
- *Taxon: Drosophila* (*Drosophila*) *nigrosparsa*
- *Sex:* both
- *Tissue:* whole body
- *Location:* Captive isofemale line Iso1 at the Molecular Ecology Group, University of Innsbruck, Austria, originally sampled at Kaserstattalm, Austria (47°7’N, 11°18’E, 2010 m a.s.l.), identified using the key in Bächli & Burla (1985).
- *Sample handling:* Individuals were removed from the stock flasks, rinsed with DMPC-treated water, placed in 1.5 ml reaction tubes, shock frozen in liquid nitrogen, and stored at -80°C. All wet-lab steps were carried out using pre-sterilized, RNase-free filter tips and tubes purchased at Corning Inc., Tewksbury, MA, USA.
- *Selection:* Poly-A(+)-mRNA
- *Layout:* Paired-end reads, 2×100 bp, 300 bp insert size
- *Library Construction Protocol:* We quantified total RNA using the Qubit RNA Assay Kit (Invitrogen, Carlsbad, CA) and used 5 μg to construct a paired-end mRNA

library. Poly-A(+)-mRNA was selected two times using Sera-Mag Oligo(dT) Magnetic Particles (Thermo Fisher Scientific, Waltham, MA). Fragmentation, reverse-transcription, end repair, A-tailing, and ligation were performed according to the instructions of the Illumina mRNA Sample Preparation Kit (Illumina, San Diego, CA). After initial size selection on an agarose gel, the library was amplified by 15 PCR cycles, followed by another purification on an agarose gel and final purification on MinElute Gel Extraction columns (Qiagen, Hilden, Germany).

Processing

Transcriptome assembly and annotation

Raw sequence data were trimmed using the Perl script trim-fastq.pl (min quality score: 20; min length: 40 bases) (http://code.google.com/p/popoolation/, Kofler *et al.* 2011) using the default settings and *de novo* assembled using Trinity (Grabherr *et al*. 2011) on a Linux machine with eight cores and 256 Gb RAM using default parameters to generate contigs as a FASTA file.

To calculate assembly statistics, we used TrinityStats (Trinity package). All assembled transcripts were annotated with BLAST+ (Camacho *et al*. 2009) by aligning them to the Flybase transcriptomes (ncRNA and CDS fasta files) of 12 *Drosophila* species (dana_r1.3; dere_r1.3; dgri_r1.3; dmel_r5.56; dmoj_r1.3; dper_r1.3; dpse_r3.1; dsec_r1.3; dsim_r1.4; dvir_r1.2; dwil_r1.3; dyak_r1.3) and to the UniProtKB/Swiss-Prot database (release 03-March-2014). To detect orthologs, we performed a reciprocal BLAST search independently with MEGABLAST and BLASTp. In this way, we were able to define homologous (with standard BLAST search), and orthologous (the reciprocal BLAST hit) transcripts (Hirsh & Fraser 2001; Jordan *et al.* 2002). Protein *i* in genome *I* is a reciprocal BLAST hit of protein *j* in genome *J*, if query of genome *J* with protein *i* yields as top hit protein *j*, and reciprocal query of genome *I* with protein *j* yields as top hit protein *i*. An *e*-value cut-off of 1e-10 was applied, and only the best hit of each query sequence was considered. For annotation, all sequences were first matched to the non-coding *Drosophila* transcripts. Then, on transcripts without matches, TransDecoder (Trinity package) was used to search for possible open reading frames (ORFs), and complete ORFs (with start and stop codons) were used in the following BLAST searches on all *Drosophila* CDS transcripts. Only the best hit of each transcript per ORF was retained and assumed as the most probable ORF for that transcript. Sequences without match were translated into amino acids and aligned to UniProtKB database with the same BLAST search strategy. For all CDS sequences where no match was found, HMMER 3.1b1 (Eddy 2011) with default settings and model-specific Pfam-A entries (Bateman *et al*. 2002) was used to search for protein domains and to confer putative functions to unknown transcripts (*e*-value cut-off 1e-10). For putative new protein-coding transcripts with at least one associated domain, the cellular localization was inferred using SignalP 4.1 server, using default parameters (Petersen *et al*. 2011), and TMHMM 2.0 server (Krogh *et al*. 2001). These algorithms predict the presence of the signal peptide cleavage sites and the location of transmembrane helices in the amino acid sequences. Where multiple ORFs for the same transcript matched Pfam domains, they were manually checked and the best ORFs selected on the base of best *e*-value or favoring ORFs with higher number of domains, since they may indicate a more organized secondary structure. Circos (Krzywinski *et al*. 2009) and the R package ggplot2 (Wickham 2009) were used to generate all graphs.

Results

Transcriptome characterization

The 140,743,918 reads (http://www.ncbi.nlm.nih.gov/sra/?term=SRS517773) were filtered (Quality scoring system phred+64, Quality scoring ASCII character range “@” to “h”) and the resulting 140,545,557 reads were assembled *de novo* into 91,048 transcripts (Dryad entry doi:10.5061/dryad.0mv56) belonging to 61,109 components (loci) with a total assembled base length of about 100 Mb. The contig N50 was of 2,247 nt with a median and average length of 493.50 nt and 1,084.85 nt, respectively. This initial assembly was annotated against over 204 k nucleotide sequences from the 12 *Drosophila* species FlyBase transcriptomes and over 542 k amino acid sequences from UniProt. More than 15 k transcripts gave significant hits with the FlyBase dataset, 44% of them had an ortholog in at least one *Drosophila* species. The match to the UniProt database gave 1,890 significant hits, of which 26% came from reciprocal BLAST search (Fig. 1A, Table 1). All transcripts with no match to database sequences and without functional domains were excluded as they may represent artifacts. The final transcriptome assembly gave a total of 18,016 transcripts (Dryad entry doi:10.5061/dryad.0mv56) belonging to 7,197 components. The total transcriptome length was 53 Mb with median and average transcript length of 2,470 nt and 2,949 nt, respectively, and mean GC content of 47.7% per transcript. There was a mean of 2.5 transcripts per component. The two most frequent numbers of transcripts per component were one (4,519) and two (1,403), while the highest number of transcripts per locus was 97. Comparing GC content and transcript lengths with the 12 known *Drosophila* transcriptomes showed no significant difference in GC percentage and a distribution of transcript lengths very similar to *D. melanogaster*, indicating a good consistency of the transcriptome reconstruction (Fig. 1B). Only 50% of nucleotides from the *D. nigrosparsa* transcripts aligned to other *Drosophila* transcriptomes (MegaBLAST search), indicating that a great portion of the transcriptome shows no similarity to other *Drosophila* species and that *D. nigrosparsa* may be rather distantly related with them. The great majority of all Flybase hits (95%) was with only three species, *D. grimshawi* (52%), *D. virilis* (38%), and *D. mojavensis* (5%), all belonging to the sub-genus *Drosophila*. Only less than 5% of the transcripts had similarities with the remaining species (Fig. 2). Functional annotations were applied using Gene Ontology (GO), revealing that more than 5 k components were associated with at least one GO term. All main three terms were covered and a good similarity to other *Drosophila* species was recovered (Fig. 3). For the more than 73 k transcripts that could not be mapped to FlyBase or UniProt database, putative new protein-coding transcripts and cellular localization were inferred by searching Pfam-A domains, signal P and transmembrane domains within their putative ORFs. HMMER found 251 transcripts belonging to 158 components with at least one Pfam domain. The two most abundant Pfam domains per component were Alcohol dehydrogenase transcription factor Myb/SANT-like (acc: PF10545.4; 10) and BTB And C-terminal Kelch (acc: PF07707.10; 7) (http://pfam.janelia.org/). Seven of the 158 components had a significant signal peptide sequence in the N-terminus. Thus, three of them had unknown function protein domains, two had enzymatic activity (Endonuclease_NS, acc: PF01223.18; Trypsin, acc: PF00089.21), one had a C-type lectin domain (acc: PF00059.16), and one had a ACP53EA domain (acc: PF06313.6), present in *Drosophila* accessory gland (seminal) proteins. Three compounds encoded putatively for proteins with transmembrane domains. The three proteins had different Pfam associated domains, one had a DM4/DM12 family domain (acc: PF07841.8), a hypothetical protein expressed in *D. melanogaster* and *Anopheles gambiae* and contained four highly conserved cysteine residues; another with a Destabilase domain (acc: PF05497.7) with lysozyme activity, and a protein with four transmembrane domains and a Pfam Tetraspannin domain (acc: PF00335.15), a protein known to act as scaffolding proteins, anchoring multiple proteins to one area of the cell membrane.

Acknowledgements

To Elisabeth Haring, for help with designing field campaigns; to Gerhard Bächli, for help with species identification; to Heike Perlinger, for assistance with fly culturing; to Clemens

Folterbauer, for assistance in the molecular laboratory. NP, VN and CS were supported by the Austrian Science Fund FWF, grant P22834. FC and FMS were supported by FWF, grant P23949.

References

Ashburner M, Bergman CM (2005) *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. Genome Research, **15**, 1661–1667.

Bächli G (2008) Drosophilidae. In: *Results from a Survey of the Biodiversity of Diptera (Insecta) in the Stilfserjoch National Park (Italy), vol. 1.* Studia Dipterolololica, **16**, 1–395.

Bächli G, Burla H (1985) Diptera drosophilidae. *Insecta Helvetica A 7*. Swiss Entomological Society Neuchâtel, Switzerland.

Bateman A, Birney E, Cerruti L *et al.* (2002) The Pfam protein families database. *Nucleic Acids Research*, **30**, 276–280.

Brown JB, Boley N, Eisman R *et al.* (2014) Diversity and dynamics of the *Drosophila* transcriptome. Nature, doi:10.1038/nature12962

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Eddy SR (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

IPCC (2013) Climate change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, New York.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, **12**, 962–968.

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Krogh A, Larsson B, von Heijne G, *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

Krzywinski M, Schein J, Birol I *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, **19**, 1639–1645.

Obbard DJ, Maclennan J, Kim KW *et al.* (2012) Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Molecular Biology and Evolution*, **29**, 3459–3473.

Petersen TN, Brunak S, von Heijne G *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, **8**, 785–786.

Pertoldi C, Bach LA (2007) Evolutionary aspects of climate-induced changes and the need for multidisciplinarity. Journal of Thermal Biology, **32**,118–124.

Wickham H (2009) ggplot2: Elegant graphics for data analysis. Springer, New York, NY, USA.

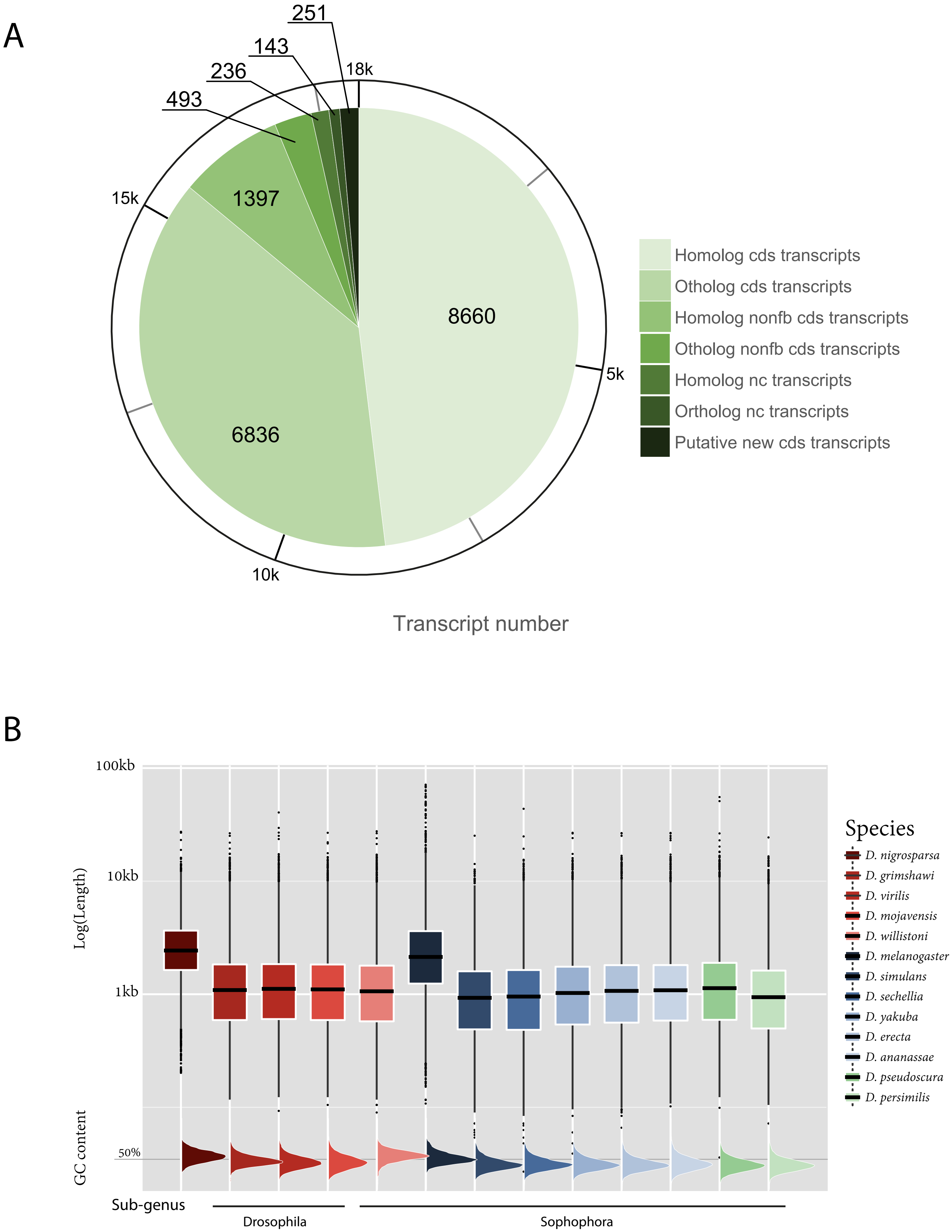
Figure 1
A) Absolute values (k = 1,000) and fraction of annotated transcripts from Flybase and UniProt databases and putative new protein-coding transcripts. Cds: protein-coding sequences; nc: non-coding; fb: FlyBase; nonfb: UniProt. B) Comparison of transcript length and GC content distributions among the 12 FlyBase *Drosophila* species with respect to all the 18,016 *D. nigrosparsa* transcripts.

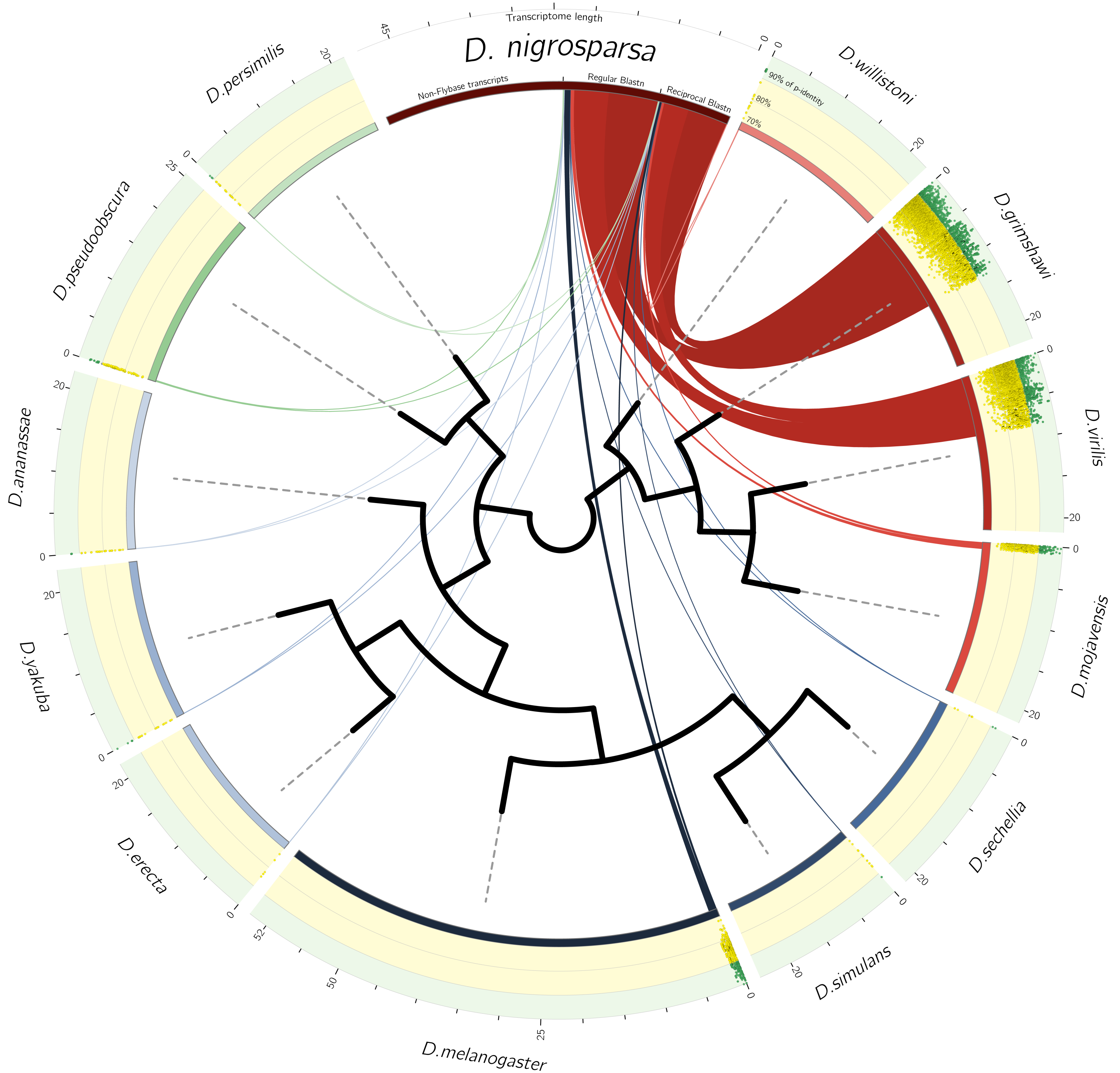
Figure 2
MegaBLAST and reciprocal BLAST best hits of *D. nigrosparsa* against 12 *Drosophila* transcriptomes. Each set of bars shows the transcriptome length in megabases for each species (bar colours are species specific). Ribbon width and colour describe the amount of aligned nucleotide length with respect to the different species; the shattered plot above bars shows the percentage of identity of each BLAST hit. Green dots are *p*-identity values higher than 90%. The most recent topology of the 12 species phylogeny (Obbard *et al.* 2012) is drawn inside the circle.

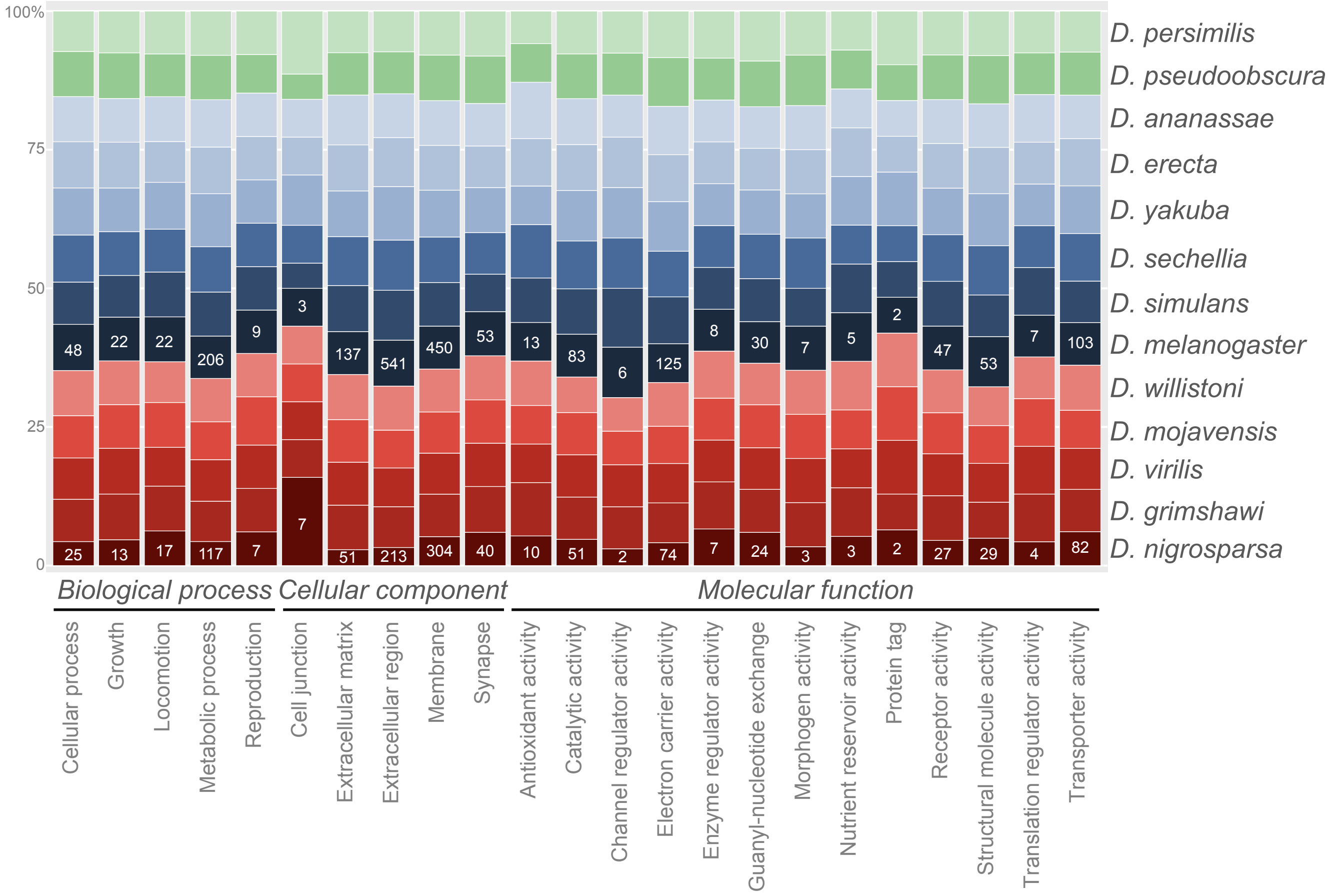
Figure 3
Stacked histogram showing the percentage of GO terms found for each transcript in the main categories. Numbers inside bars are the absolute numbers of terms in *D. nigrosparsa* (bottom) and *D. melanogaster*, the latter being used as reference to evaluate how each category is represented in *D. nigrosparsa* annotation.

Table 1: Run, alignment, and annotation statistics

paired-end reads	140,743,918
paired-end reads after quality filtering	140,545,557
Trinity transcripts	91,048
Total annotated transcripts	18,016
Annotated protein coding transcripts	17,637
FlyBase hits	15,875
UniProt hits	1,890
New putative CDS	251


A
251
143
236
493
18k
15k
5k
10k
1397
8660
6836
Homolog cds transcripts
Otholog cds transcripts
Homolog nonfb cds transcripts
Otholog nonfb cds transcripts
Homolog nc transcripts
Ortholog nc transcripts
Putative new cds transcripts
Transcript number
B
100kb
10kb
1kb
Log(Length)
GC content
50%
Species
D. nigrosparsa
D. grimshawi
D. virilis
D. mojavensis
D. willistoni
D. melanogaster
D. simulans
D. sechellia
D. yakuba
D. erecta
D. ananassae
D. pseudoscura
D. persimilis
Sub-genus
Drosophila
Sophophora


D. nigrosparsa
Transcriptome length
Non-Flybase transcripts
Regular Blastn
Reciprocal Blastn
90% of p-identity
80%
70%
D.willistoni
D.grimshawi
D.virilis
D.mojavensis
D.sechellia
D.simulans
D.melanogaster
D.erecta
D.yakuba
D.ananassae
D.pseudoobscura
D.persimilis
0
20
25
45
50
52


100%
75
50
25
0
D. persimilis
D. pseudoobscura
D. ananassae
D. erecta
D. yakuba
D. sechellia
D. simulans
D. melanogaster
D. willistoni
D. mojavensis
D. virilis
D. grimshawi
D. nigrosparsa
48
22
22
206
9
3
137
541
450
53
13
83
6
125
8
30
7
5
2
47
53
7
103
25
13
17
117
7
7
51
213
304
40
10
51
2
74
7
24
3
3
2
27
29
4
82
Biological process
Cellular component
Molecular function
Cellular process
Growth
Locomotion
Metabolic process
Reproduction
Cell junction
Extracellular matrix
Extracellular region
Membrane
Synapse
Antioxidant activity
Catalytic activity
Channel regulator activity
Electron carrier activity
Enzyme regulator activity
Guanyl-nucleotide exchange
Morphogen activity
Nutrient reservoir activity
Protein tag
Receptor activity
Structural molecule activity
Translation regulator activity
Transporter activity