Title

Transcriptomic resources for three populations of *Conus miliaris* (Mollusca: Conidae) from Easter Island, American Samoa and Guam

Authors

David A. Weese[a,b*]

Thomas F. Duda[a,c]

[a]Department of Ecological and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI, 48108, USA

[b]Department of Biological and Environmental Sciences, Georgia College and State University, Milledgeville, GA, 31061, USA

[c]Smithsonian Tropical Research Institute, Balboa, Ancón, 0843-03092, Republic of Panama

* Corresponding author (david.weese@gcsu.edu)

Introduction

Species interactions represent fundamental ecological processes that can have significant impacts on the evolutionary trajectories of species. However, the contribution of predator-prey interactions to genetic and phenotypic divergence within and between species remains largely unknown. In this context, predatory marine snails belonging to the family Conidae exhibit considerable variation in venom composition, a phenomenon that may be due to the evolution of venom components (e.g., 'conotoxins' and other 'conopeptides' used in the venom to capture prey) in response to predator-prey interactions (Duda et al. 2009). It has been hypothesized that geographic differences in prey utilization drive the evolution of venom components in cone snails and that the diversity of conotoxins is positively associated with prey diversity (Duda and Palumbi 1999). Given that each *Conus* species can possess a repertoire of 50-200 different venom components (Olivera 2006; Hu et al. 2011; Lluisma et al. 2012), to test this hypothesis requires broad, yet detailed, surveys of the diversity of genes expressed in the venom and their expression patterns. In this context, high-throughput RNA-sequencing offers an opportunity to survey a large variety of expressed genes including those encoding venom components. Additionally, this approach can facilitate the development of genetic resources for organisms that have been traditionally understudied such as members of the Conidae.

Conus miliaris is a broadly distributed cone snail found in tropical and subtropical waters of the Indo-West Pacific from the Red Sea and eastern shores of Africa to Easter Island and Sala y Gómez islands in the western Pacific (Röckel et al. 1995). Throughout much of its range, *C. miliaris* co-occurs with as many as 36 congeners (Kohn 2001) and preys almost exclusively on three species of eunicid polychaetes (Kohn 1978). However, *C. miliaris* is the only member of Conidae with an established population on Easter Island where it occurs at higher densities than at other localities throughout its range (Kohn 1978). Additionally, the diet of *C. miliaris* at Easter Island is considerably broader than elsewhere in its range and includes additional eunicids as well as species of nereids, an onuphid and members of other polychaete families (Kohn 1978). Given the high population density, lack of congeners and relatively broad diet of *C. miliaris* at Easter Island compared to other localities, it has been suggested that *C. miliaris* has undergone ecological release at Easter Island. This ecological release experienced by *C. miliaris* at Easter Island offers a unique opportunity to investigate the effects of predator-prey interactions on the evolution of toxin components. Based on previous studies of patterns of variation at two conotoxin loci and one mitochondrial gene (cytochrome oxidase I) (Duda and Lee 2009a; 2009b), *C. miliaris* at Easter Island is genetically differentiated from other populations in the Indo-West Pacific and harbors a distinct repertoire of conotoxins. Moreover, the conotoxin loci appear to have been subject to directional or disruptive selection likely associated with the increased dietary breadth of the Easter Island population (Duda and Lee 2009a). Nonetheless, these studies were limited to only two toxin loci belonging to the same superfamily and one mitochondrial gene. Additional nuclear and conotoxin loci are needed to more thoroughly investigate possible associations between diets and venoms. Thus, developing transcriptomic resources in the form of RNA-sequencing data for *C. miliaris* from Easter Island and other Indo-West Pacific populations will allow comparative analyses of conotoxin gene diversity and expression patterns to be conducted while providing valuable genomic information for other lines of study as well.

To determine if differences in dietary breadth are associated with 1) difference in venom composition and/or 2) variation in conotoxin expression levels among geographic populations of *C. miliaris*, this study details the assembly, annotation and broad scale comparisons of transcriptomes from three geographic populations of *C. miliaris* from Easter Island, Guam and American Samoa. In addition to these transcriptomic resources, this study also identifies 1000's of single nucleotide polymorhphisms (SNPs) suitable for population demography, structure and connectivity analyses. Although a number of recent studies have utilized high-throughput sequencing technologies to describe the venom duct transcriptomes of *Conus* species (e.g., Hu et al. 2011; Hu et al. 2012; Lluisma et al. 2012; Terrat et al. 2012; Lu et al. 2014), the majority of studies have centered around the potential for neurobiological and therapeutic applications and the data from these studies are not typically publicly available (for exception see Hu et al. 2012). Additionally, as of July 2014, this study represents the first population level analyses of venom duct transcriptomes from any venomous organism (i.e., snakes, spiders, scorpions, etc.) and represents novel genetic resources that should prove useful to the larger scientific community.

Data Access

- NGS sequence data: Available from NCBI SRA under PRJNA257931, SRP045405.
- Assembled transcripts: Available from Dryad for *C. miliaris* as a whole as well as for each geographic population as FASTA files under accession doi: 10.5061/dryad.t74q4.
- Relative expression levels: Available from Dryad as tab-delimited files corresponding to the number of reads mapping back to each annotated transcript of the *C. miliaris* assembly and FPKM values for each of the 22 individuals, geographic populations as well as for *C. miliaris* under accession doi: 10.5061/dryad.t74q4.
- Putative single nucleotide polymorphisms (SNPs): Available from Dryad as a VCF file under accession doi10.5061/dryad.t74q4.

Meta-information

- Sequencing center: University of Michigan DNA Sequencing Core, University of Michigan Medical School, Ann Arbor, MI 48108.
- Platform and model: Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA, USA).
- Design description: Goals were to 1) generate novel transcriptomic data for *Conus miliaris* as a whole and for three geographic populations; 2) annotate transcripts for each population and the species as a whole; 3) determine which genes were among the most highly expressed for each population; 4) identify and determine the relative abundance of conotoxin related transcripts, and; 5) identify potentially informative single-nucleotide-polymorphic (SNPs) markers for population genetic studies.
- Analyze type: cDNA.

- Run date: 31 January 2014.

Library

- Strategy: non-normalized RNA.
- Taxa: *Conus miliaris (*Hwass in Bruguière, 1792).
- Sex: The sex of each individual snail is as follows: CmilEI120, male ; CmilEI165, female; CmilEI202, female; CmilEI283, female; CmilEI329, female; CmilEI381, male; CmilEI388, male; CmilEI400, male; CmilGU075, male; CmilGU078, female; CmilGU085, male; CmilGU096, female; CmilGU100, male; CmilGU104, female; CmilGU117, female; CmilGU139, male; CmilAS014, female; CmilAS027, male; CmilAS028, male; CmilAS032, male; CmilAS035, male; and CmilAS038, female.
- Tissue: Dissected whole venom ducts for all individuals.
- Location: Specimens of *C. miliaris* were collected at three localities in the Indo-West Pacific: 1) Hanga Roa, Easter Island (27.1°S, 109.4°W) in November 2007, 2) Pago Bay, Guam (13.4°N, 144.8°E) in June 2008, and 3) Fatumafuti, American Samoa (14.9°S, 170.7°W) in March 2010.
- Sample handling: At each site, collected snails were kept in ~100 ml of seawater in individual containers (3 ounce plastic cups). Once snails defected, feces were preserved for diet analysis and venom ducts were dissected and preserved in RNAlater (Qiagen, Valencia, CA, USA) and stored at –80 °C.
- Selection: Total RNA.
- Layout: 100 bp paired-end reads.
- Library construction protocol: After being removed from RNAlater (Qiagen, Valencia, CA, USA), RNA was extracted by homogenizing whole venom ducts in Trizol (Invitrogen, Carlsbad, CA, USA) and purified following the manufacturer's instructions. Genomic DNA was then removed from the total RNA using a RapidOut DNA Removal Kit (Thermo Fisher Scientific, Waltham, MA, USA). The quality of the RNA was assessed with a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Approximately 100 ng/µl of total RNA per sample were submitted to University of Michigan DNA Sequencing Core (https://victor.brcf.med.umich.edu/wiki/index.php) for library preparation and indexing using the Illumina Tru-Seq© Kit. A single flowcell lane on an Illumina HiSeq 2000 was used to sequence all 22 cDNA libraries, yielding a total of 349,293,902 reads with an average of 15,876,996 (± 2,052,016 reads) reads per library/individual. The number of reads per geographic population were 131,525,916, 130,566,380 and 87,201,606 for Easter Island, Guam and American Samoa respectively. The overall quality of the data was high as the % of bases having a quality score ≥ 30 and the Mean Quality Score of the reads averaged 92.56% and 35.93, respectively, across all

individuals. Raw reads for each individual are available from NCBI's SRA under BioProject PRJNA257931 and SRP accession number SRP045405.

Processing

- Preprocessing of reads: Given the overall high quality of the reads (see above) and the fact that aggressive trimming based on quality scores can negatively impact assembly of RNA-Seq data (MacManes 2014; D.A. Weese personal observation), reads were not filtered or quality trimmed prior to assembly. However, for expression estimation and SNP identification, reads were trimmed based on a phred-scale quality score cut-off of 20 with reads shorter than 20 bases being discarded using the fastx toolkit (http://hannonlab.schl.edu/fastx_toolkit/).

- Transcript assemblies: For the total *C. miliaris* dataset, 100 bp PE reads for all 22 individuals were combined and, to improve computational performance (Hass et al. 2013), were digitally normalized using `normalize-by-median.py` (Brown et al. 2012). The transcriptome was then assembled *de novo* from the digitally normalized reads using default parameters and 12 CPU cores (--CPU 12) in Trinity (version Trinityrnaseq_r20140413p1; Grabherr et al. 2011). For the Easter Island, Guam and American Samoa datasets, reads for individuals were combined based on geographic location and assembled using Trinity and the settings described above. While the species assembly took ~22 hours to complete and resulted in 204,951 transcripts, population assemblies took < 12 hours to complete and resulted in an average of 127,717 transcripts per transcriptome. Transcript characteristics for *C. miliaris* and each population are presented in Table 1. All transcript assemblies are available as FASTA files from Dryad under accession doi: 10.5061/dryad.t74q4.

- Annotation of transcripts: To annotate the assembled transcripts for each dataset (whole species and three geographic populations), assemblies were first compared to a reference database of non-redundant conotoxin and conotoxin signal sequences downloaded from the Conoserver database (Kass et al. 2008; http://research1t.imbuq.edu.au/conoserver/) using BLASTx (version 2.2.29; Altschul et al. 1997) and retaining matches with an *e*-value < 10^{-5}. Transcripts that failed to match any sequences in our conotoxin database were then compared to local databases comprised of annotated proteins from the two mollusc species available on the NCBI Unigene database, *Aplysia californica* and *Lottia gigantea* (downloaded from http://www.ncbi.nlm.nih.gov/unigene), all *Conus* related proteins downloaded from NCBI's nr database (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/) as well as the Swiss-Prot (downloaded from http://www.uniprot/) database using the BLASTx algorithm. In these cases, NCBI's default parameters were used and hits with an *e*-value < 10^{-5} were retained. Parsed BLAST reports for all transcripts as well as only the toxin related transcripts can be accessed on Dryad (doi: 10.5061/dryad.t74q4).

- Relative expression levels: To generate an estimate of the number and identity of genes being expressed at relatively high levels for each individual and population, trimmed single-end (SE) reads for each individual as well as each population were mapped onto the *C. miliaris* annotated transcriptome using bowtie v1.0.1 (Langmead et al. 2009).

Then, the perl script `align_and_estimate_abundance.pl` (included in the Trinity package) was used to estimate expression vales (*i.e.*, raw counts and FPKM values) for each individual and/or population. Annotated tab-delimited count files with the number of reads mapping back to each transcript for all transcripts as well as toxin related only transcripts for all individuals and populations are available from Dryad under accession doi: 10.5061/dryad.t74q4.

- Identification of putative single nucleotide polymorphisms (SNPs): To identify potential SNPs, trimmed paired-end (PE) reads for each individual were remapped to the *C. miliaris* annotated transcriptome (serving as the index) using Bowtie 2 v2.1.0 (Langmead and Salzberg 2012) in –local mode. Resulting SAM (Sequence Alignment/Map) files were converted to binary format (i.e., BAM), sorted and read group information was added using SAMtools v0.0.19 (Li et al. 2009). Afterwards, duplicate reads were marked using the MarkDuplicates utility of Picard (http://picard.sourceforge.net). Following this, the Genome Analysis Toolkit (McKenna et al. 2010; DePristo et al. 2011) was used for SNP detection. First, individual BAM files were realigned around indels and variants were called with the HaplotypeCaller using the suggested setting for RNA-seq data (-recoverDanglingHeads, -dontUseSoftClippedBases, -strand_call conf 20.0, -stand_emit_conf 20.0). Variants for each sample were then annotated and filtered with the VariantAnnotator and VariantFiltration tools using the filters for RNA-seq data (i.e., -window 35, -cluster 3, FS > 30.0, QD > 2.0). Lastly, all samples were jointly genotyped using the GenotypeGVCFs tool. The resulting VCF file can be accessed on Dryad under accession doi: 10.5061/dryad.t74q4.

Results

- A greater number of individuals (8) were sequenced from the Easter Island and Guam populations than from the American Samoa population (6). Given this, it is not surprising that a larger number of transcripts were assembled for the Easter Island and Guam populations ($\bar{x}$ =131,046 transcripts) than for the American Samoa population (87,201,606 transcripts).

- Annotation statistics for each transcriptomes are presented in Table 1. Annotation of the *C. miliaris* assembly resulted in ~ 15 % of the assembled transcripts having similarity to proteins in our protein databases. Of the 31,099 transcripts that could be annotated, 516 (1.7%) were identified as being related to venom components. Annotation of the population assemblies resulted in ~1/6 ($\bar{x}$ =21,808 transcripts) of transcripts per transcriptome having BLASTx matches to our protein databases. For any population, only a small percentage (<2%) of annotated transcripts were identified via BLASTx searches as being toxin related. In general, as the number of transcripts in the assemblies decreased, the percentage of transcripts that could be annotated increased (e.g., >18% of transcripts in the American Samoa assembly could be annotated). This general trend likely reflects the complexity of the assemblies; having more individuals included in an assembly increases the complexity of the assembly and increases the likelihood of miss-assembled transcripts (i.e., artifacts of the assembly process). Despite the fact that a large number of transcripts remain unannotated, the number of uniquely annotated transcripts

(unique hits from the BLASTx searches) for the *C. miliaris* assembly (25,131) is in line with the number of genes that have been described for the molluscs *Lottia gigantea* (23,851, Lotgi v1.0), *Haliotis rufescens* (22,000; De Wit and Palumbi 2013) and *Pinctada fucata* (23,257; Takeuchi et al. 2012).

- Statistics for mapping single-end trimmed reads to the annotated transcripts of the *C. miliaris* Trinity assembly are presented in Table 2. Of the 174,554,450 *C. miliaris* reads, 39% (68,641,977) were mapped successfully to the annotated transcripts. For the population level analyses, an average of 39% of reads mapped back to the annotated reference transcripts with an average of 887, 993 and 674 reads per transcript for Easter Island, Guam and American Samoa respectively. For all datasets, mapping of reads to the annotated *C. miliaris* transcripts resulted in a bimodal coverage distribution, with a large number of transcripts having either fewer than ten (45%) or more than a thousand (15%) reads (Figure 2). Not surprisingly, in all cases the vast majority of mapped reads mapped to toxin-related transcripts. For example, >70% of mapped reads in the Guam dataset mapped to toxin-related transcripts with an average of ~42,000 reads per transcript. When the top 100 of highly expressed transcripts are examined on a per dataset basis, ~60-75% had a BLASTx “hit” to annotated conotoxin proteins in the Conoserver and NCBI's nr databases (Table 3).

- After stringent filtering as recommended by the Broad Institute, 585,736 high-quality SNPs were identified from the *C. miliaris* individuals. Of these, 20,957 occurred within toxin-related transcripts and represent good candidate markers for exploring the role of selection on conotoxins. The large number of SNPs identified in this study should provide a valuable resource for future population demographic, structure and connectivity studies of *Conus* miliaris and other closely related species.

- Quality scoring system: Phred+33.

- Quality scoring ASCII character range: "!” to "J".

Acknowledgements

Funding from the National Geographic Committee for Research and Exploration (CRE 8228-07) and the National Science Foundation (IOS 0718370) supported collection of specimens used in this work. An Associate Professor Support Funds award from the University of Michigan's College of Literature, Science and the Arts supported the sequencing work and analyses. Permits for sample collection at Easter Island were obtained with the assistance of Javier Rivera Vergara, Brian Dyer and Liliana Cortés from the Unidad de Recursos Bentónicos of the Subsecretaria de Pesca of Chile. We would like to than Dan Chang for assistance with field collections at Easter Island. Barry Smith, Alex Kerr, Jason Biggs and others at the University of Guam Marine Lab coordinated field work in Guam and University of Guam students Marielle Terbio, Chris Rosario, Cabrinie Rivera and Jonathan Lim provided lab and field assistance.

Literature cited

Altschul SF, Madden TL, Schaffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25,** 3389–3402.

Brown CT, Howe A, Zhang Q, *et al.* (2012) A Reference-free algorithm for computational normalization of shotgun sequencing data. Available: arXiv:1203.4802

De Pristo M, Banks E, Poplin R *et al. (*2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491-498.

De Wit P, Palumbi SR (2013) Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation. *Molecular Ecology*, **22**, 2884-2897.

Duda TF Jr., Palumbi SR (1999) Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 6820-6823.

Duda TF Jr., Chang D, Lewis BD, Lee T (2009) Geographic variation in venom allelic composition and diets of the widespread predatory marine gastropod *Conus ebraeus*. *PloS One*, **4**(7), e6245.

Duda TF Jr., Lee T (2009a) Ecological release and venom evolution of a predatroy marine snail at Easter Island. *PloS One*, **4**(5), e5558.

Duda TF Jr., Lee T (2009b) Isolation and population divergence of a widespread Indo-West Pacific marine gastropod at Easter Island. *Marine Biology,* **156**, 1193-1202.

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

Hass BJ, Papanicolaou A, Yassour M *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols,* **8,** 1494–1512.

Hu H, Bandyopadhyay PK, Olivera BM *et al.* (2011) Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. *BMC Genomics*, **12**, 60.

Hu H, Bandyopadhyay PK, Olivera BM *et al.* (2012) Elucidation of the molecular evenomation strategy of the cone snail *Conus geographus* through transcriptome sequencing of its venom duct. *BMC Genomics,* **13**, 284.

Kaas Q, Westermann JC, Halai R *et al.* (2008) ConoServer , a database for conopeptide sequences and structures. *Bioinformatics*, **3**, 445-446.

Kohn AJ (1978) Ecological shift and release in an isolated population – *Conus miliaris* at Easter Island. *Ecological Monographs,* **48**, 323–336.

Kohn AJ (2001) Maximal species richness in Conus: diversity, diet and habitat on reefs of northeast Papua New Guinea. *Coral Reefs,* **20**, 25–38.

Langmead B, Trapnell C, Pop M *et al.* (2009) Ultrafast and memory-effecient alignment of short DNA sequences to the human genome. *Genome Biology,* **10**:R25.

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357-359.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2079-2079.

Lluisma AO, Milash BA, Moore B *et al.* (2012) Novel venom peptides from the cone snail *Conus pulicarius* discovered through next-generation sequencing of its venom duct transcriptome. *Marine Genomics*, **5,** 43-51.

Lu A, Yang L, Xu S *et al.* (2014) Various conotoxin diversifications revealed by a venomic study of *Conus flavidus*. *Molecular & Cellular Proteomics*, **13**, 105-118.

MacManes MD (2014) On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics,* **5,** 1–7.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

Olivera BM (2006) Conus peptides: biodiversity-based discovery and exogenomics. *Journal of Biological Chemistry*, **281**, 31173-31177.

Röckel D, Korn W, Kohn AJ (1995) Manual of the living Conidae. Verlag Christa Hemmen. Wiesbaden, Germany.

Takeuchi T, Kawashima K, Koyanagi R, *et al.*, (2012) Draft genome of the pearl oyster *Pinctada fucata*: A platform for understanding bivalve biology. *DNA Research*, dss005.

Terrat Y, Biass D, Dutertre S *et al.* (2012) High-resolution picture of a venom gland transcriptome: Case study with the marine snail *Conus consors*. *Toxicon,* **59**, 34-46.

Table 1: Summary for the assembly and annotation of the *C. miliaris*, Easter Island, Guam and American Samoa transcriptomes.

	C. miliaris	**Easter Island**	**Guam**	**American Samoa**
Assembly/Transcript Characteristics				
Total number of reads	349,293,902	131,525,916	130,566,380	87,201,606
Total number of transcripts	204,951	149,583	134,741	98,828
Number of total bp in transcripts	1227,925,804	83,827,208	82,227,639	42,388,575
Average transcript length (bp)	624.2	560.4	550.5	428.9
Maximum transcript length (bp)	25,902	23,251	16,001	15,460
N50 total transcript length (bp)	879	719	693	463
N40 total transcript length (bp)	1,181	962	922	594
N30 total transcript length (bp)	1,588	1,303	1,250	791
N20 total transcript length (bp)	2,174	1,807	1,737	119
N10 total transcript length (bp)	3,179	2,664	2,595	1,748
Transcript Annotation				
Number of transcripts with BLASTx hit	31,099	23,166	24,303	17,956
% of transcripts with BLASTx hit	15.2	15.5	16.3	18.2
Number of toxin related transcripts	516	435	340	284
% of annotated transcripts related to toxins	1.7	1.9	1.4	1.6

Table 2: Summary of mapping statistics for each set of reads to the annotated transcripts of *C. miliaris*.

	C. miliaris	**Easter Island**	**Guam**	**American Samoa**
Number of trimmed SE reads	174,554,450	65,725,283	65,247,848	43,581,319
Annotated transcripts				
Number of mapped reads	68,641,977	24,142,467	27,897,897	16,601,613
% of reads mapped	39	37	43	38
Average fragments per transcripts	2,235.05	886.79	993.46	674.09
Minimum fragments per transcript	1	1	1	1
Maximum fragments per transcript	6,927,047	4,735,273	2,363,361	1,515,817
Average FPKM	49.16	52.47	55.6	61.73
Minimum FPKM	0.01	0.01	0.01	0.01
Maximum FPKM	162,007	180,902	242,859	88,260
Toxin related transcripts				
Number of mapped reads	45,136,449	14,277,864	20,098,939	10,760,812
% of mapped reads	66	59	72	65
Average fragments per transcripts	88,502.84	29,807.65	41,875.74	24,512.10
Minimum fragments per transcript	1	1	1	1
Maximum fragments per transcript	6,927,047	4,735,273	2,363,361	1,515,817
Average FPKM	2374.77	2,182.41	2,783.43	2,813.09
Minimum FPKM	0.04	0.04	0.08	0.08
Maximum FPKM	162,007.29	180,901.66	242,858.65	88,260.11

Table 3. Summary of functional annotation based on BLASTx searches for the top 100 most highly expressed transcripts for the *C. miliaris*, Easter Island, Guam and American Samoa datasets.

	C. miliaris	**Easter Island**	**Guam**	**American Samoa**
Conotoxin	74	58	56	77
Ribosomal proteins	1	11	11	0
Mitocondrial proteins	5	3	6	4
Elongation factors	2	4	4	2
Actin/myosin	4	4	4	2
Arginine kinase	1	1	1	0
Hypothetical/uncharacteristized proteins	2	2	4	3
Other characterized proteins	11	17	14	11
Total	100	100	100	100

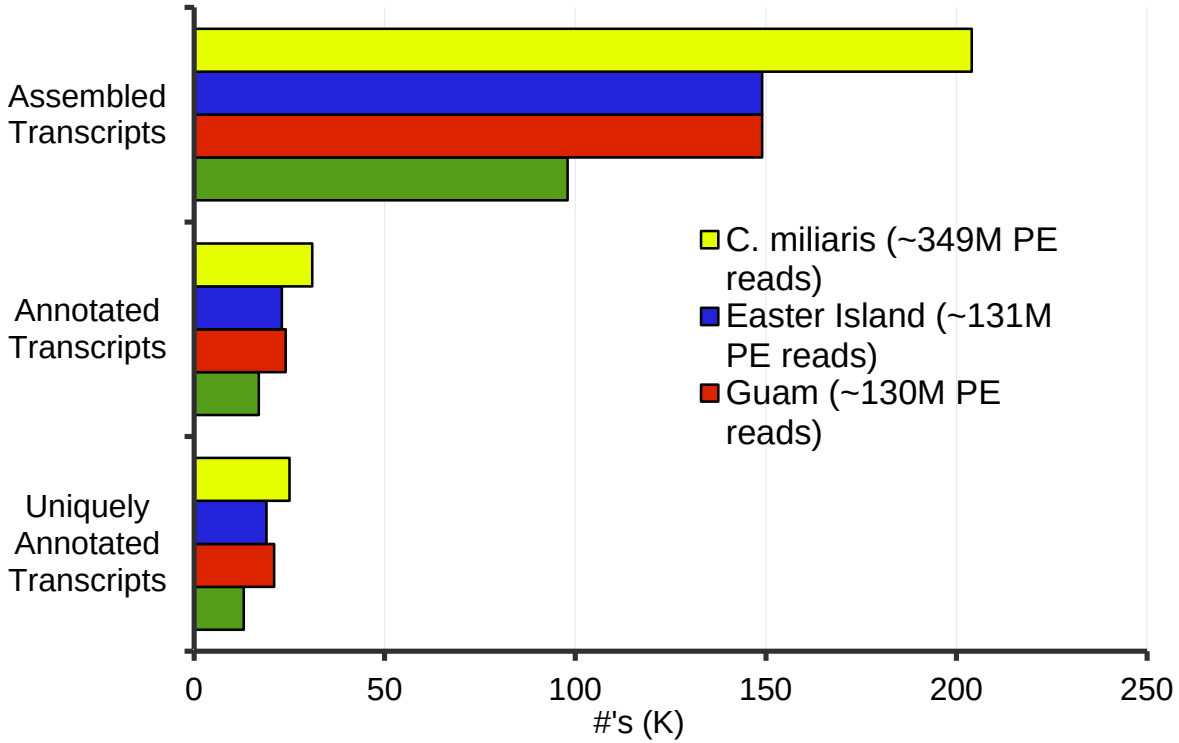

Figure 1. Summary statistics for transcriptome assemblies and their annotation for *C. miliaris* and each population (see text for additional details).

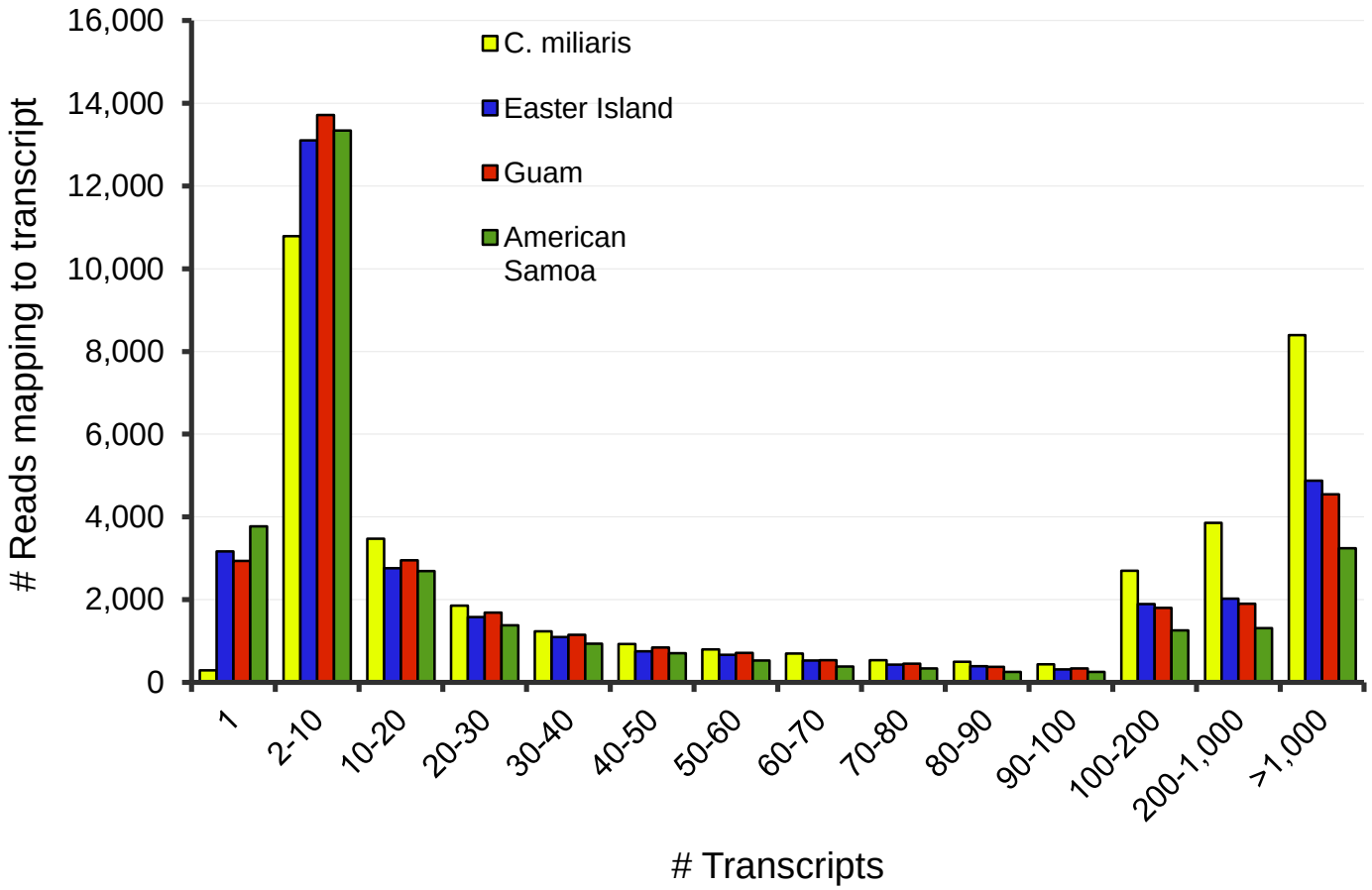

Figure 2. Histogram showing frequency of annotated Trinity transcripts with total numbers of reads mapping back to those transcripts for each assembly.