Bowhead Whale Genome Project: Progress on the Transcriptome.

John W. Bickham[1], Gary W. Stuart[2], John C. Patton[1], John C. George[3], and Robert S. Suydam[3].
[1]Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907;
[2]Department of Biology, Indiana State University, Terre Haute, IN 47809; [3]North Slope Borough,
Department of Wildlife Management, Barrow, AK 99723, USA

Abstract

Genome sequencing of non-model species previously has been used to study aspects of genome regulation, structure and evolution including characterization of expressed sequence tags, identification of xenobiotics, documenting gene ontologies, genetic diversity, and estimating relative levels of gene expression within and among tissues (Hale et al., 2010). To begin characterization of the bowhead whale genome, RNA was extracted from heart, cerebellum, liver, and testes from bowhead number 10B16, and from the retina of bowhead 10B20. Using the Illumina HiScan next generation sequencing platform at Purdue's genome sequencing core center we obtained 13,162,565,851 bases of sequence which passed quality filter. The Trinity *de novo* assembly software successfully assembled 423,657 contigs from those data, including 157,699 large contigs (>500 base pairs). There were 81,319 contigs annotated by the Blast2Go software based on the possession of open reading frames identified in the assembled transcripts. Messages homologous to about 60% of the known human genes were identified in the bowhead transcriptome (i.e., about 14,000 genes identified) using the Bowtie and Tophat software.

Introduction

The bowhead whale (*Balaena mysticetus*) is remarkable for being the second largest animal by mass (>100 tons), the oldest living animal (>200 years), and remarkably healthy with few reported diseases or cancer. It was hunted to near extinction during the commercial whaling era and presently it is an important food resource for native villages in Alaska and Russia. Two of the four recognized stocks of bowhead whales were hunted to very low levels and have not recovered; the Sptizbergen stock is extinct or nearly so, and the Sea of Okhotsk stock numbers only a few hundred individuals. However, the eastern Canadian Arctic population numbers about 8,000 thousand (Heide-Jorgensen et al. 2006) and the Bering-Chukhi-Beaufort Seas (BCB) stock has recovered from a low of about 1,000 to about 13,000 and continues to grow at a rate of 3.4% per year (George et al., 2004). It is the BCB stock that is hunted by the Russian and Alaskan native communities and is the subject of this study.

The International Whaling Commission (IWC) provides a quota for the aboriginal hunt of bowheads, and genetics issues including stock structure, diversity levels, gene flow estimates, genetic bottleneck effects, to name a few, are important considerations in the establishment of the quota. Therefore, the development of methods to study the genetic variation of this species has importance to the ongoing management efforts of the IWC and the Alaska Eskimo Whaling Commission.

The rapid advancement of next-generation DNA sequencing technology is profoundly changing the field of genetics. Natural resource and conservation genetics has until now largely focused on a few useful genetic markers including mitochondrial DNA, nuclear microsatellites, SNP loci, and some targeted nuclear loci. But the field is in a transitional period in which the application of genomics methods will play an ever increasing role. The genome sequences of relatively few wildlife species have been completed, and among cetaceans this includes only *Tursiops truncatus* which is still in the process of being compiled (http://www.ensembl.org/Tursiops_truncatus/Info/Index?db=core).

Population genomics, the large scale comparison of DNA sequences (Black et al., 2001), has already proven useful to the understanding of evolutionary history, forensics, and medicine (Jorde et al., 2001). Crawford and Lazzaro (2012), using simulation methods, found that estimates of population genetic differentiation and population growth parameters were systematically biased when inference was based on 4x sequencing, but biases were markedly reduced at 8x and higher read depth. This indicates that methods to produce and analyze wide-scale genomic comparisons among individuals and populations are nearly feasible. This paper presents the results of an analysis of the bowhead whale transcriptome based upon a comparison of five tissues and two individuals. The purpose of the study is to develop a database of genomic sequence to serve as a resource to develop genome-wide variability assays suitable for answering the key issues related to stock structure, evolution, management, and the basic biology of this key indicator species of the Arctic environment.

Materials and Methods

RNA isolation and cDNA construction.—Tissue biopsies were obtained from two male bowhead whales harvested by Eskimo hunters at Barrow, Alaska during the Fall hunt of 2010; heart, cerebellum, liver, and testes were biopsied from bowhead number 10B16, and retina from bowhead 10B20. Samples were immediately placed in liquid nitrogen and transported in a dry shipper to Purdue University. RNA was extracted using TRIZOL reagent (Invitrogen) with methods following the manufacturer's protocol. RNA was purified using a Invitrogen PureLink Micro-to-Midi columns from the Total RNA Purification System using the standard protocol. RNA quantity and quality was estimated with a spectrophotometer (Nanodrop) and by gel electrophoresis using an Agilent model 2100 Bioanalyzer. cDNA libraries were constructed by random priming of chemically sheared poly A captured RNA. Randomly primed DNA products were blunt ended . Products from 450-650 base pairs were then isolated using a PippenPrep. After the addition of an adenine to the fragments a Y primer amplification was used to produce properly tailed products.

Sequencing and assembly.-- . Paired-end sequences of 100 base pairs per end were generated using the Illumina HiScan platform.. Sequences with primer concatamers, weak signal, and/or poly A/T tails were culled. The Trinity software package for *de novo* assembly (Grabherr et al., 2011) was used for transcript reconstruction. Assemblies were annotated using Blast2Go software with homology settings of e $\leq$ 1.00 X 10^{-03} and bit scores $\leq$ 40. The Blast2Go software identifies probable homologies based on the possession of open reading frames identified in the assembled transcripts. All sequences with significant BLASTx hits were noted, as was the species which produced the top BLAST hit. Gene function was also inferred using Blast2Go software which searches for Gene Ontology (GO) terms (Ashburner et al. 2000; Shaw et al. 1999). The same settings as the BLASTx search were used to confirm significant ontology which was categorized as Molecular Function, Biological Process, or Cellular Component.

Single nucleotide polymorphism (SNP) identification.—Only SNPs identified with $Q \geq 20$ and $DP \geq 5$ were used (240,173, including duplicates in alternative transcripts). Heterozygous SNPs (hSNPs) were called using the mpileup program in SAMtools (Li et al., 2009) using the default $GQ \geq 3$.

Results

A total of 138,495,774 sequence reads comprising >13 billion bp remained after quality control and primer trimming. The numbers and sizes of reads and contigs are reported in Table 1. The total number of annotated contigs was 81,319. However, this is likely inflated greatly by "broken" contigs (incomplete assembly of transcripts). By grouping identical annotations, this number is reduced to about 23,000. However, this is still likely an over estimate. Using the Bowtie and Tophat programs with the human transcriptome data, the estimated number of bowhead contigs identified as being homologous to human genes was approximately 14,000 or ca. 60% of the known human genes.

The results of Blast searches to identify homologies in the GenBank database are shown in Figures 1 and 2. Figure 1 shows the species distribution of total blast hits, which is predominated by identities to the human genome. This is due to more thorough coverage and annotation of the human genome and transcriptome relative to all other vertebrates. The rank of species homology to the bowhead transcriptome with this method does not reflect taxonomic relationships but is driven primarily by biased sequence representation in the databases. Figure 2 shows the single top hit (i.e., the greatest sequence identity to bowheads) and this metric more closely tracks phylogenetic relationship with the cow (Bos taurus) having the highest number of top hits. This is the species most closely related to bowheads for which a genome sequence has been produced and deposited in GenBank.

Table 2 shows the estimated frequencies of SNPs among the 5 tissues sampled. The two individuals sampled can be compared by reference to retina (bowhead 10B20) and Tissues 1-4 (bowhead 10B16). The data are shown for 8 size classes of contigs. As contig size increases, the frequency of estimated SNPs increases. With this method, there appears to be approximately 0.5-0.6 SNPs per 1,000 bases of poly A selected RNA per individual whale.

Discussion

Genome studies of non-model organisms are made difficult by the lack of reference genome with which to compare sequences. This study demonstrates the ability of the Illumina next generation sequencing platform to produce an extensive representation of the transcriptome of a non-model species, such as the bowhead whale. This method produces relatively short (100 bp) reads from each end of any RNA (cDNA) fragment, nonetheless it has the advantage over certain other next-generation sequencing methods in providing considerably more reads per run. In this case, we produced > 13 billion bases of quality sequence which was successfully organized into 157,699 large contigs. Significantly, our initial BLAST analysis yielded >81,000 annotated contigs and over 60 percent of the known human genes were recovered in this initial sequencing effort. Therefore, the data successfully captures ample sequence homology to provide identification of a substantial fraction of the RNA messages produced among the 5 tissues studied. Our studies will now focus on using the transcriptome database for the development of SNP analysis methods for population genetics studies, and further exploring the transcriptome to better understand the basic biology and health status of the bowhead whale.

Acknowledgments

We thank the Alaska Eskimo Whaling Commission (AEWC) and the Barrow Whaling Captains' Association for their confidence, guidance and support of our research. We gratefully acknowledge funding provided by the North Slope Borough Department of Wildlife Management and National Oceanic and Atmospheric Administration (through the AEWC).

Literature Cited

Ashburner MC, Ball JA, Blake A et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29.

Black IV WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. Annual Review of Entomology **46**, 441-469.

Crawford JE, Lazzaro BP (2012) Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Frontiers in genetics, **2012(66),** 1664-8021.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE **6**(5): e19379. doi:10.1371/journal.pone.0019379

George JC, Follman E, Zeh J, Sousa M, Tarpley R, Suydam R (2004) Inferences from bowhead whale corpora data, age estimates, length at sexual maturity and ovulation rates. Paper SC/56/BRG8 presented to the IWC Scientific Committee, June 2004. Available from The International Whaling Commission, The Red House, 135 Station Road, Impington, Cambridge, Cambridgeshire CB24 9NP, United Kingdom.

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

Hale MC, Jackson JR, DeWoody JA (2010) Discovery and evaluation of candidate sex-determining genes and xenobiotics in the gonads of lake sturgeon (*Acipenser fulvescens*). Genetica **138**,745–756.

Heide-Jorgensen MP, Laidre KL, Jensen MV, Dueck L, Postma LD (2006) Dissolving stock discreteness with satellite tracking: Bowhead whales in Baffin Bay. *Marine Mammal Science* **22**, 34-45.

Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics* **10**, 2199-2207.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Margulies MM, Egholm W, Altman E et al (2005) Genome sequencing in microfabricated high-density pico litre reactors. *Nature* **437**, 376–380.

Shaw DR, Ashburner M, Blake JA et al. (1999) Gene Ontology: a controlled vocabulary to describe the function, biological process and cellular location of gene products in genome databases. American Journal of Human Genetics **65,** 1419.

Table 1.—Results of RNA sequencing of 5 tissues from two bowhead whales. All Reads refers to all sequenced fragments of any size, Large Contigs includes contigs comprised of multiple reads of 500 bp or larger, and All Contigs refers to small and large contigs combined. All values are in bp.

All Reads	**Total reads**	138,495,774
	Total bases	13,162,565,851
	Size range of reads	2-101
	N50 (modal size)	101
	Average length	95
Large Contigs	**Contig size**	≥500
	Total large contigs	157,699
	Total number of bases	322,342,312
	Contig size range	500-24765
	N50 (modal size)	3,442
	Average length	2,044
All Contigs	**Total number of contigs**	423,657
	Total number of bases	401,340,157
	Contig size range	201-24765
	N50 (modal size)	2,436
	Average length	947
Annotations	**Number of annotated contigs**	81,319

Table 2.—SNP frequencies estimated for each tissue per size class of contigs. Tissues 1-4 are from bowhead 10B16 and retina is from 10B20.

Contig Size (bp)	Tissue					
	1. Cerebellum	2. Heart	3. Liver	4. Testes	5. Retina	Tissues 1-4
≥201	2.7E-04	2.7E-04	2.7E-04	2.8E-04	3.1E-04	3.9E-04
>500	3.3E-04	3.2E-04	3.2E-04	3.4E-04	3.8E-04	4.8E-04
>1000	3.6E-04	3.5E-04	3.6E-04	3.8E-04	4.2E-04	5.2E-04
>2000	3.9E-04	3.8E-04	3.8E-04	4.1E-04	4.5E-04	5.6E-04
>3000	4.0E-04	3.9E-04	3.9E-04	4.2E-04	4.6E-04	5.7E-04
>4000	4.2E-04	4.0E-04	4.0E-04	4.4E-04	4.7E-04	5.9E-04
>5000	4.3E-04	4.0E-04	4.0E-04	4.5E-04	4.7E-04	6.0E-04
>6000	4.5E-04	4.2E-04	4.2E-04	4.7E-04	4.9E-04	6.2E-04

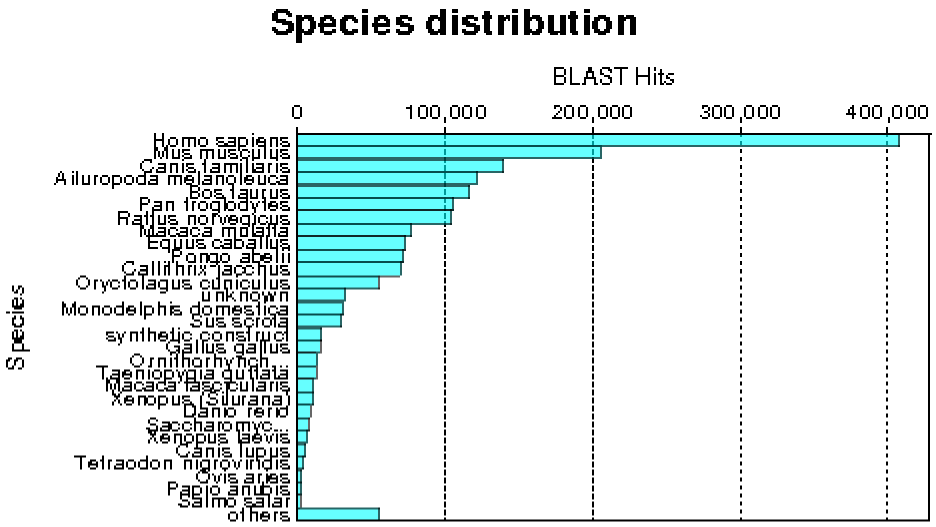

Figure 1.—The species distribution of total BLAST hits of reads from the bowhead whale transcriptome is primarily driven by man and mouse, which have the most complete genome sequence analyzed.

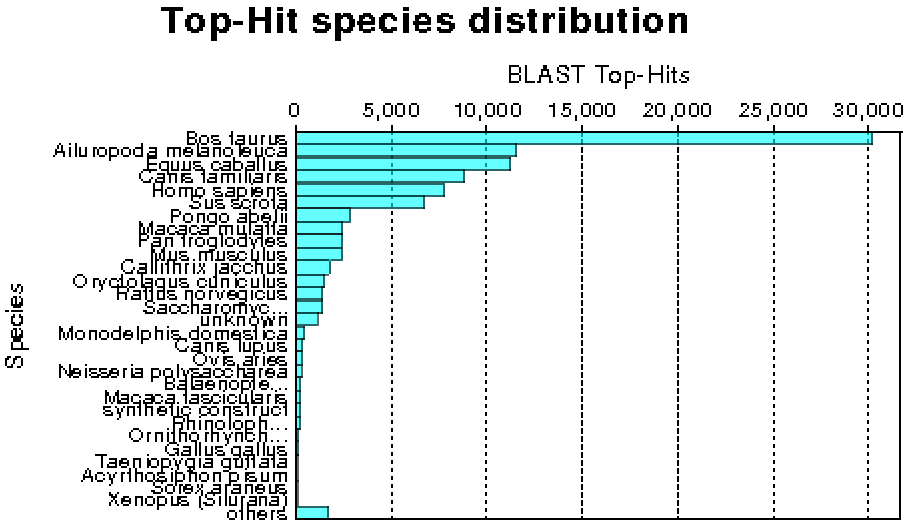

Figure 2.—The species distribution of the top Blast hits of reads from the bowhead whale transcriptome show the greatest homology to the cow genome, which is phylogenetically the closest relative of the bowhead with a completed genome sequence.