

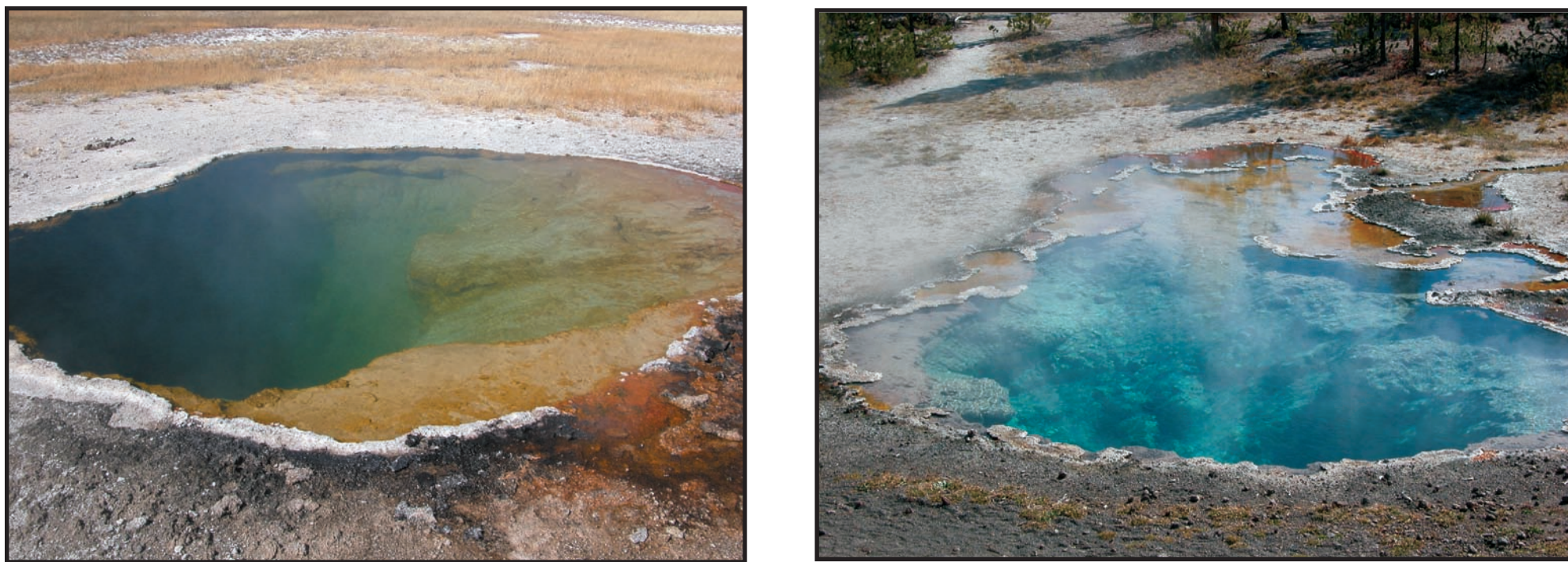
Assembly of Viral Metagenomes from Yellowstone Hot Springs

Tom Schoenfeld¹, Melodee Patterson¹, Paul M. Richardson², Eric Wommack³, Mark Young⁴, David Mead¹

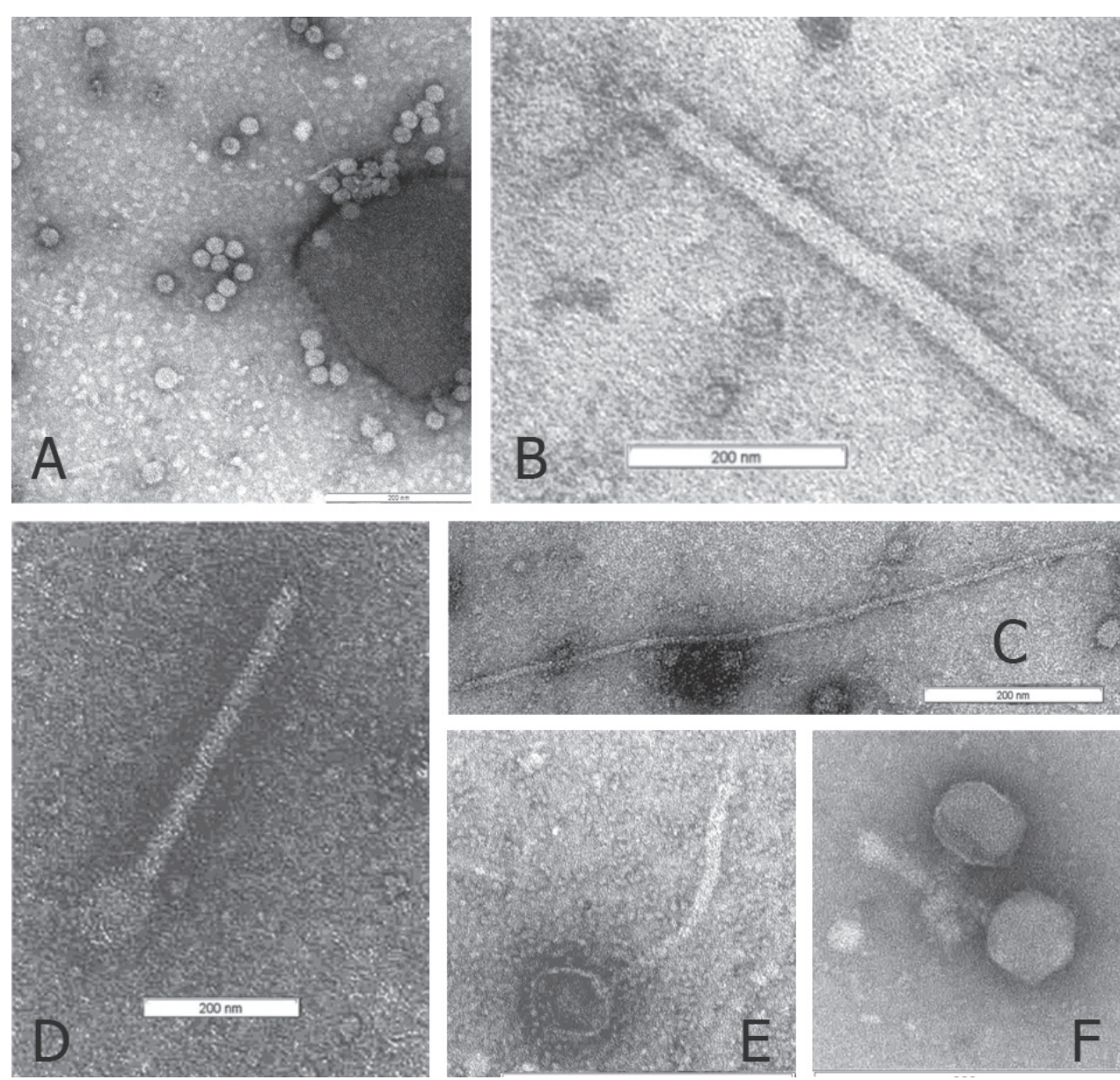
¹Lucigen Corporation, Middleton, WI, ²Department of Energy Joint Genome Institute, Walnut Creek, CA, ³Department of Plant and Soil Sciences, University of Delaware, Newark, DE, ⁴Plant Sciences & Plant Pathology, Montana State University, Bozeman, MT

Abstract

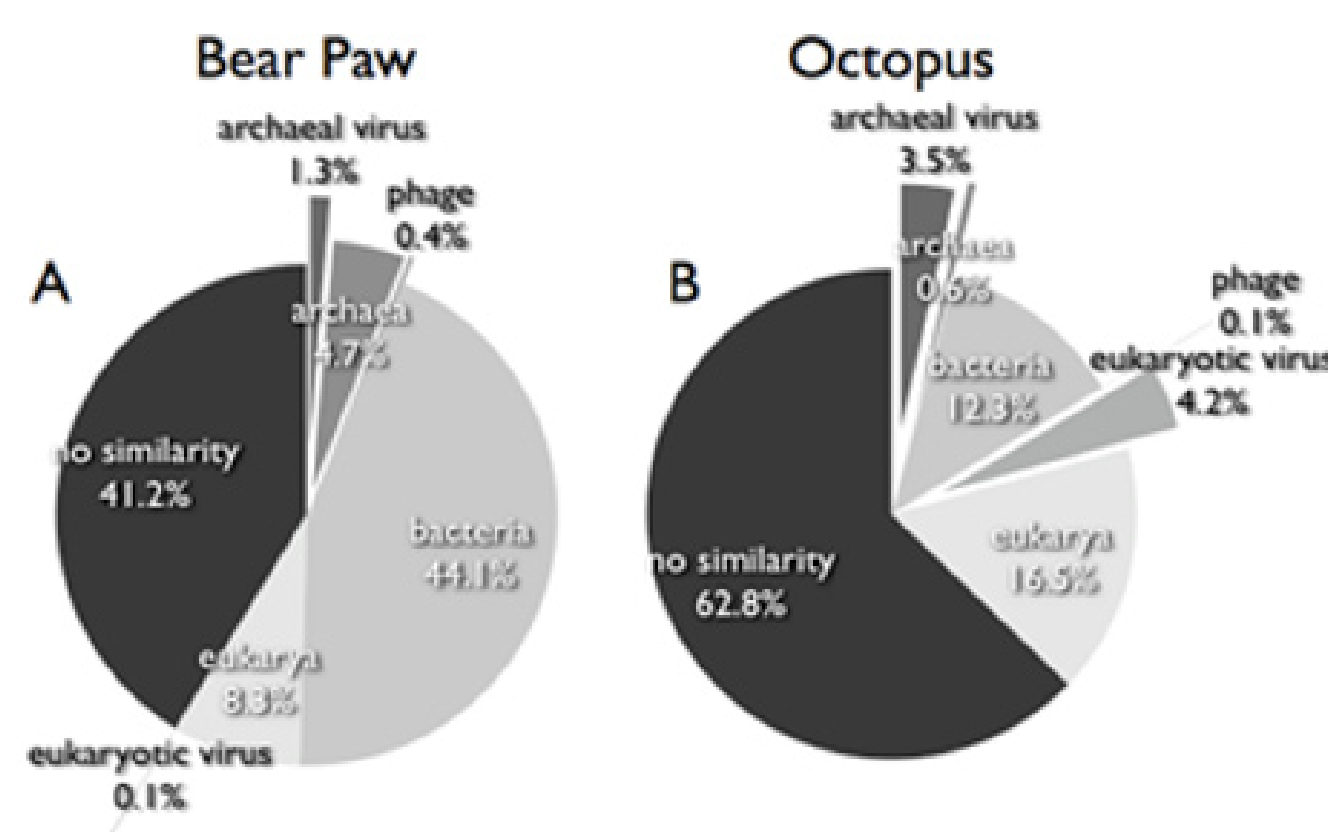
Thermophilic viruses were first reported decades ago; however, knowledge of their diversity, biology and ecological impact is limited. Previous research on thermophilic viruses has focused on cultivated strains. This study examined metagenomic profiles of viruses directly isolated from 74° to 93°C mildly alkaline hot springs. Viral abundance ranged from 10e5 to 10e6 per ml. Using a new method for constructing libraries from picogram amounts of DNA, nearly 30 Mb of viral DNA sequence from two hot springs was determined. Approximately 25% of the viral sequences share regions of significant similarity with the other hot spring. Although most sequences were unrelated to known genomes, hundreds of BLASTx similarities provide insights into viral lifestyles in this environment. In contrast to previous viral metagenomic studies, sequences were assembled at 50% identity, creating composite contigs as large as 35 kb that show the inherent heterogeneity in the populations. One 16.5-kb composite contig encodes 26 apparent virus-associated genes, including three clones that express functional DNA polymerases. Lowering assembly identity to 50% from the standard 95% reduced the number of different viral types to 300 from 1400. The 50% assembly included one contig of high similarity and perfect synteny to nine genes from *Pyrobaculum* spherical virus (PSV), a cultured thermophilic crenarchaeal virus. In fact, nearly all the genes of the 28-kb genome of PSV have apparent homologs in the metagenomes. Similarities to thermoacidophilic viruses isolated on other continents were limited to specific open reading frames but were equally strong. Metagenomics provides a powerful tool to study the diversity of viruses in these extreme environments.



Bear Paw Hot Spring (74°C) left and Octopus Hot Spring (93°C) right



TEM Images of virus-like particles directly isolated from YNP hot springs.

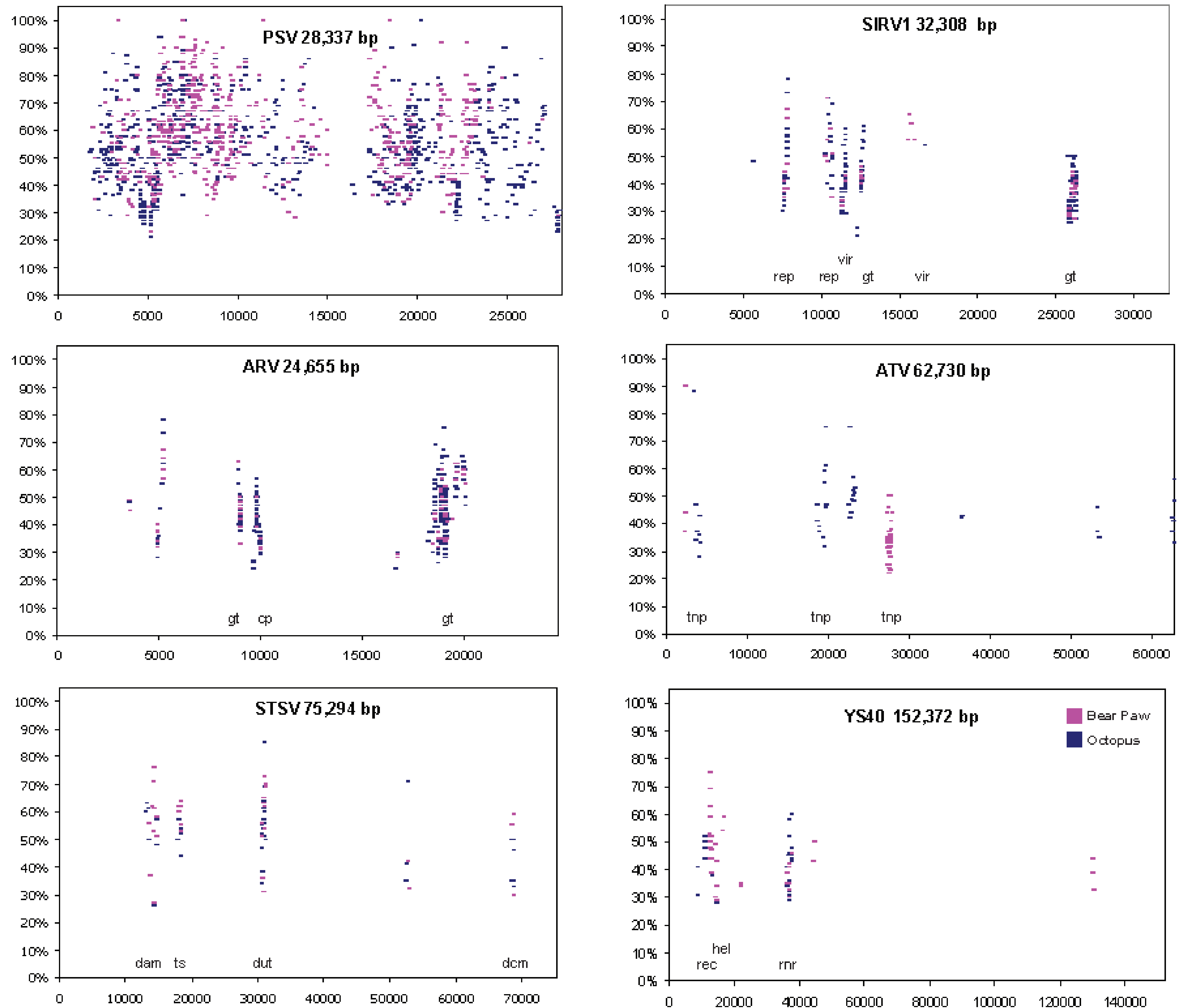


Broad classification of viral metagenomic contigs based on tBLASTx similarities. Contigs assembled at 95% identity from Bear Paw and Octopus reads (Panel A and B, respectively) were compared to sequences in GenBank to infer phylogeny. Shown are frequencies of contigs with no significant sequence similarity in GenBank ($E < 0.001$) and those with sequence similarity to Bacteria, Archaea, Eukarya and their respective viruses.

Numbers of tBLASTx similarities to cultivated viruses in metagenomic libraries.

Virus (Acc. No.)	Bearpaw	Octopus
ARV, <i>Acidilobus</i> rod-shaped virus	36	228
SIRV, <i>Sulfolobus islandicus</i> rod-shaped virus	30	217
PSV, <i>Pyrobaculum</i> spherical virus	44	152
SIFV, <i>S. islandicus</i> filamentous virus	7	46
STSV1, <i>Sulfolobus tengchongensis</i> spindle-shaped virus 1	26	22
ATV, <i>Acidilobus</i> two-tailed virus	8	17
TTSV1, <i>Thermoproteus tenax</i> spherical virus 1	6	12
YS40, <i>Thermus thermophilus</i> YS40 phage	15	41
Twort, <i>Staphylococcus</i> phage Twort	4	21

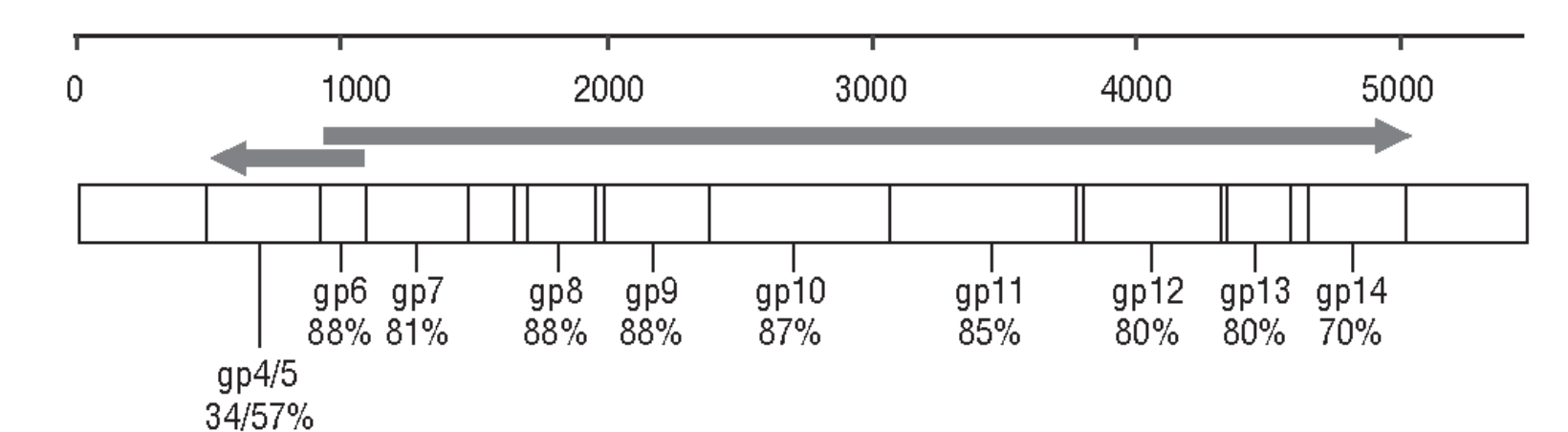
Alignment of Octopus and Bear Paw viral metagenomic library contigs with cultured virus genomes.



Contigs from the viral metagenomic libraries were compared by tBLASTx to the genomes of PSV, SIRV1, ARV, ATV, STSV and YS40. Each bar represents a unique alignment of the metagenomic sequence to the indicated location on the cultivated viral genome, shown on the horizontal axis. Percent coding sequence identities are shown in the vertical axis. Red bars indicate Bear Paw alignments; blue bars indicate Octopus alignments. Also shown are the known or predicted functions of the conserved coding sequences (*rep*, replication related, *vir*, virion component; *gt*, glycosyltransferase; *tnp*, transposase; *cp*, coat protein; *dam*, adenine DNA methylase; *ts*, thymidylate synthase; *dut*, dUTPase; *dcm*, cytosine DNA methylase; *hel*, helicase; *rec*, recombinase; *mrr*, ribonucleotide reductase).

Sequence assembly data and estimation of viral diversity.

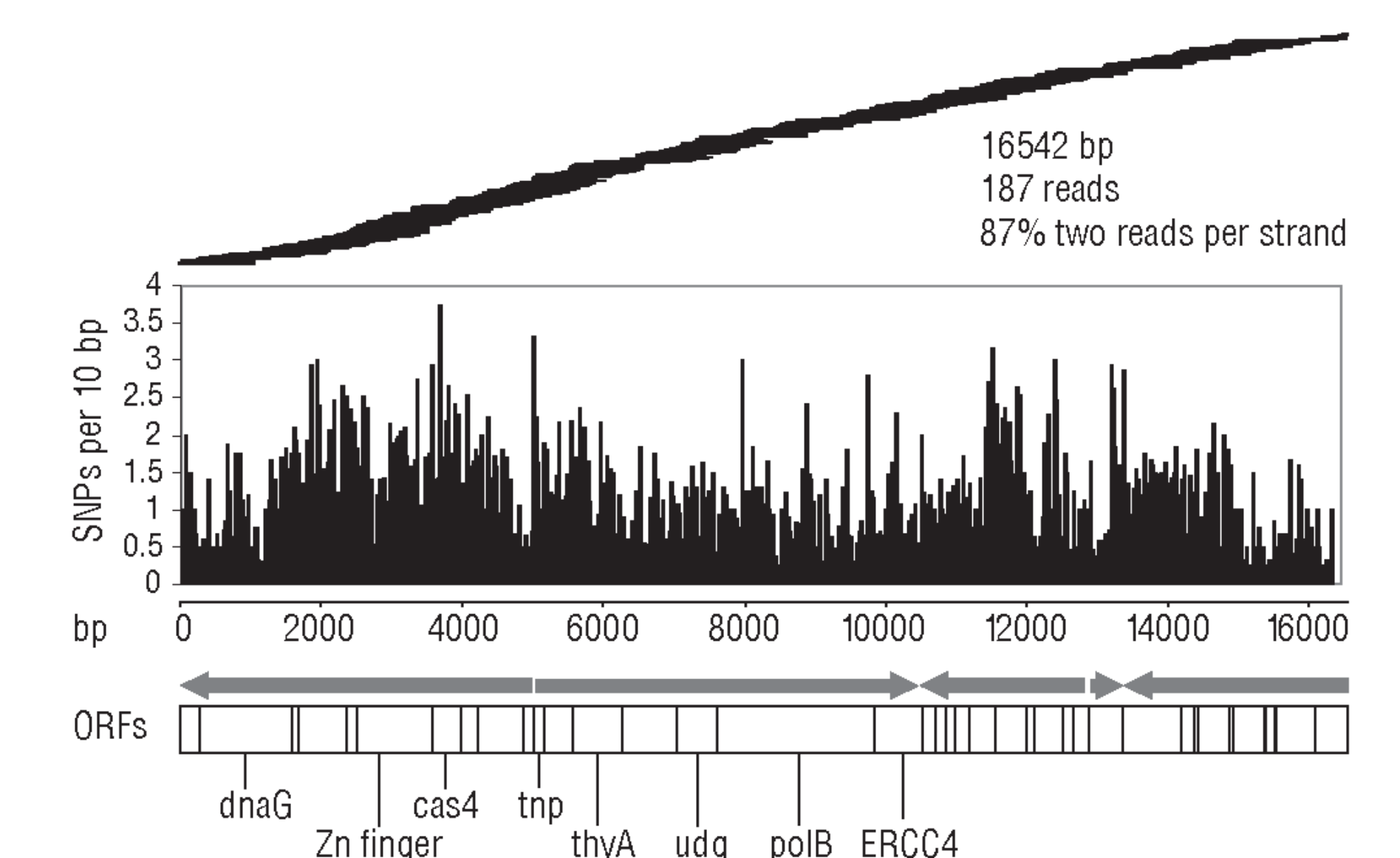
	Bearpaw	Octopus	Totals
Sequence reads	7,685	21,196	28,883
	Bear Paw 95%	Octopus 95%	Bear Paw 50% Octopus 50%
Contigs assembled	6,191	13,543	4,850
Avg. reads per contig	1,239	3,129	1,587
Largest contig (nt)	3,503	4,554	8,007
Power law richness	1,440	1,310	548
Evenness score	0.946	0.954	0.933
Most abundant virus	2.14 %	1.88 %	3.93 %
Shannon-Wiener score	6.88	6.85	5.88



Genes and gene order are highly conserved between *Pyrobaculum* spherical virus and a consensus contig from the Bear Paw library. Contig 372 (5492 bp, 71 reads) was assembled at $\geq 50\%$ identity from the Bear Paw library. Open reading frames identified by GeneMark algorithm were compared by BLASTp to proteins in GenBank. Similarities to *Pyrobaculum* spherical virus proteins are shown with percent coding identity. The gene names are based on the annotation in GenBank and are named in order of their location on the viral chromosome. Direction of transcription is indicated by the arrows.

Largest contigs from 50% assemblies.

Contig	Length	Total Seq Length	No. of Seq.	Average coverage	Cumulative length
Octopus					
1496	35089	1814653	1601	51.72	35089
591	18974	1421079	1115	74.9	54063
1596	18451	360922	349	19.56	72514
426	17755	453426	414	25.54	90269
722	16542	182178	187	11.01	106811
1539	16465	383575	369	23.3	123276
Bear Paw					
480	8007	43088	47	5.38	8007
687	6276	75303	81	12	14283
1151	5749	44544	49	7.75	20032
372	5492	66055	71	12.03	25524
793	5219	27159	29	5.2	30743
754	5027	21341	23	4.25	35770



Alignment of nucleotide polymorphisms with coding sequences in a 16.5 kb consensus contig from Octopus Hot Spring. Contig 722 was assembled at $\geq 50\%$ identity from the Octopus library. Sequence coverage is shown on the top, with each line representing a separate read. Single nucleotide polymorphisms per 10 base pairs were normalized to the number of reads covering the respective nucleotide (middle) and are aligned with predicted open reading frames from the consensus sequence in the contig (bottom). Direction of transcription is shown by the arrows. Similarities to known genes were identified by BLASTp.