Supplemental Material

**The draft genome assembly of *Dermatophagoides pteronyssinus* supports identification of novel allergen isoforms in Dermatophagoides species**

Thomas A. Randall1, Jim Mullikin2, NISC Comparative Sequencing Program2,

Geoffrey A. Mueller1\*

Short Title: Draft Genome of *Dermatophagoides pteronyssinus*.

1 Intramural Program National Instititue of Environmental Health Sciences

2 NIH Intramural Sequencing Center, National Human Genome Research Institute

\* Corresponding author

Geoffrey.Mueller@nih.gov

111 T.W. Alexander Drive, MD-MR01

Research Triangle Park, NC, 27709 USA

**Methods**

DP was cultured at room temperature on Diet A as described [1]. As food became depleted, mites of all active life stages migrated onto the lids, were collected by aspiration onto a 38 mm stainless steel and were killed by freezing. Once food was completely consumed, the spent culture was also killed by freezing. Both mites and spent culture were lyophilized prior to use. Two aqueous mite extracts were prepared from the mite material described above using methods adapted from [1]. Both extracts were prepared in sterile Dulbecco’s Phosphate Buffered Saline. One extract was prepared in the presence of HALT Protease Inhibitor Cocktail and EDTA (Thermo Scientific) according to the manufacturer’s instructions. The second extract was prepared without these inhibitors. Mites were collected as indicated above from culture lids and surface sterilized as described for scabies mites [2]. Genomic DNA was isolated from mites of all active stages using the Wizard SV genomic DNA purification system. To enable proper digestion of mite tissues using this kit, mites were ground in digestion buffer using a Dounce homogenizer.  RNA was isolated from surface-sterilized mites using the Direct-zol RNA MiniPrep w/ TRI-Reagent The sample prep was similar to that for DNA with ~50 mg of live mites being ground in 500 ul of TriReagent on ice.

**Genome Assembly**

The genome assembled from 303,594 PacBio ccs reads generated by the NIH Intramural Sequencing Center (NISC) using SPAdes 3.8.1 [3]. SPAdes assembles genomes using a range of kmers and the K127 genome was the most complete, as judged by having the fewest scaffolds and the highest N50 and was used as the starting point for the final assembly. From the initial assembly of 2486 scaffolds, those under 300 bp were removed, resulting in a preliminary assembly of 879 scaffolds. Separately, 216,079,207 Illumina 126 bp paired end reads were assembled with Phusion [4], generating an approximately 70 Mb assembly with an N50 of 8.7 kb. BLAST of both assemblies to an NCBI bacterial genome collection showed that the predominant component of the microbiome was found to be *Serratia marcescens*. The genome of *S. marcescens* WW4 (Genbank accession CP003959.1) was found to be the top hit. 55 contigs with an E value of <= 1-30 were removed as potential bacterial contamination to result in a final assembly of 834 contigs with an N50 of 376 kb and a genome size 52.5 kb with a GC content of 29.03 %. Repeat content was determined using misa for microsatellites (default settings) [5] and Repeatmasker [6].Dotplot alignments were performed with Gepard 1.4 (PMID: [**17309896**](http://www.ncbi.nlm.nih.gov/pubmed/17309896)**).** All raw genomic data used in this study has been submitted to either SRA as Bioproject PRJNA395246.

**Protein prediction and comparison**

Whole organism RNA-seq data for DP was obtained as described [7]. 317,387,496 300 bp paired-end reads were assembled using Trinity [8] and from this, Transdecoder predicted a set of 25,446 proteins from which a non-redundant set of 14,409 proteins > 100 aa was obtained. To explore the genome for proteins that may have not been identified in our RNA-seq sample, we used two algorithms, Augustus 2.2.5 [9] and SNAP (snap-2013-11-29) [10], to independently predict proteins sets from our genome assembly. Of the available training sets for each, we empirically determined that the jewel wasp *Nasonia vitripennis* training set was the most optimal available for protein prediction. Augustus predicted 9006 proteins; 3536 not in the RNA-seq dataset while SNAP predicted 17,876 proteins; 6480 not in the RNA-seq dataset. Combined, these two algorithms predicted a total of 4959 proteins > 100 aa not previously predicted. The final set contains 19,368 proteins (Supplemental file 1). We also used SNAP (same training set as above) to make protein predictions for the five species in Table 1 for which there were no publicly available predicted proteomes and other *Sarcoptes scabiei* (SS) genomes. We used BUSCO v1.1b1 (ref; using the arthropod geneset) along with the extended CEGMA dataset of 2748 genes to assess the completeness of our genome. For genome-level orthology prediction of proteomes of the six mites and ticks for which we had publicly available protein predictions, we used Orthovenn, with an E value of 1e-05 and an inflation of 1.5 [11]. For orthology prediction of specific classes of gene sets of interest, we used the script orthoparahomlist.pl with a cutoff E = 1e-05 [2].

**Results**

**Mitochondial DNA**

The mitochondrial (mt) genome of DP has previously been published, and is approximately 14.2 kb in size [12]. A full length mt genome of 14,680 bp was extracted from the PacBio data, this was 95.17 % identical with, and co-linear with the published DP mt genome (Supplemental Figure 2) and contains the same complement of 13 CDS and 22 tRNAs. The 12S ribosomal RNA in the our mt genome has a four bp deletion and the 16S ribosomal gene has 7 SNPs and a single base pair deletion, both relative to the published genome. The major difference between the two mt genomes is a largely AT rich region of 495 bp (Supplemental Figure 3). This insertion occurs within the D loop located at 11483-11768 of the published mitochondrial genome. The D loop itself is a variable region itself involved in both replication and the transcription of nearby genes [13].

**Microbiome**

The composition of some of the microbiome could be assessed because DNA was extracted from whole mites. The Phusion genome assembly was examined using a collection of complete bacterial genomes downloaded from NCBI for better quantification. Any contig with an E value of 1e-30 or lower was considered to be bacterial contamination. Supplemental Table 4 and Supplemental Figure 5 shows that the microbiome of DP is dominated by *Serratia marcescens* but a range of Bartonella species were also present. The PacBio assembly was able to generate a nearly complete 5 Mb genome of *S. marcensens*; this was most closely related to *S. marcescens* WW4. The microbiome has a very distinct GC content from the *D. pteronyssinus* genome. For example, the reconstructed *S. marcescens* genome is 59 % GC, consistent with published *S. marcescens* genomes, in contrast to 29 % for the DP genome. For comparison, the DF genome is 29.5 % GC and the SS genome is 33.26 % GC. A dotplot alignment of these is shown in Supplemental Figure 4. In contrast, the microbiome of DF was predominantly Enterobacter species [14]. However, we note that the Wizard SV genomic DNA purification system used for DP was not optimized for lysis of Gram + bacteria, so these results should be not be considered as a direct analysis of the DP microbiome. For more information, a comprehensive review of the microbiomes of 143 mite species was recently published [15]. Consistent with our data, this review shows that the microbiome of mites that have been examined are dominated by gram negative Enterobacteria.

**Repeat Content**

The repeat content of the DP genome was examined using several algorithms. Misa was previously used to characterize the simple sequence repeat (SSR) content of SS [2,16,17], which, along with DF was shown to have a much elevated level of SSRs (>2400 per Mb) compared to the other sequenced mite genomes. Our analysis of SSRs within DP with Misa supported this as the total number of identified SSRs is 103,106 and we see a level of just under 2000 SSR per Mb supporting the previous suggestion that the Hypoorder of Astigmata is selecting for an elevated level of SSRs. RepeatMasker was used to assess the general repeat content of *D. pteronyssinus*. The results of both these analyses are in Supplemental Table 5.

**Transfer RNA**

We used tRNAscan-SE (Supplemental Table 5) to identify the tRNAs within our assembly [18]. We found 68 tRNAs, a similar complement to that of DF (65 tRNAs) which we confirmed by re-running the tRNAscan analysis on the DF genome. However, the genomic organization of the tRNAs may be quite different between the two species. In DF, 26 of the tRNAs are located within 185 kb on a single genomic scaffold (KN266225.1) while the tRNA complement of DP is much more widely dispersed; only one scaffold contained five tRNAs. A caveat to this interpretation is that a different fragmentation of our assembly could, in part, cause this difference; however, our genome N50 of 376 kb compares favorably with that of the DF genome (187 kb). The exon/intron structure of the tRNA genes of these species differ also in that 33 of the DF tRNAs, including all 26 on scaffold KN266225.1, are intronless while only 18 of the 68 tRNAs in DP do not contain an intron.

**References**

1 Avula-Poola S, Morgan MS, Arlian LG: Diet Influences Growth Rates and Allergen and Endotoxin Contents of Cultured Dermatophagoides farinae and Dermatophagoides pteronyssinus House Dust Mites. Int Arch Allergy Imm 2012;159:226-234.

2 Rider SD, Morgan MS, Arlian LG: Draft genome of the scabies mite. Parasite Vector 2015;8:585.

3 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA: SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19:455-477.

4 Mullikin JC, Ning ZM: The phusion assembler. Genome Res 2003;13:81-90.

5 Thiel T, Michalek W, Varshney RK, Graner A: Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 2003;106:411-422.

6 Smit AFA, Hubley R, Green P: RepeatMasker,

7 Ogburn RN, Randall TA, Xu Y, Roberts JH, B. M, Karnuta JM, Rider SD, Kissling GE, London RE, Pomes A, Arlian LG, Fitzgerald MC, Mueller GA: Are dust mite allergens more abundant and/or more stable than other Dermatophagoides pteronyssinus proteins? J Allergy Clin Immun 2017

8 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;29:644-652.

9 Stanke M, Waack S: Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 2003;19:Ii215-Ii225.

10 Korf I: Gene finding in novel genomes. BMC bioinformatics 2004;5

11 Wang Y, Coleman-Derr D, Chen GP, Gu YQ: OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research 2015;43:W78-W84.

12 Dermauw W, Van Leeuwen T, Vanholme B, Tirry L: The complete mitochondrial genome of the house dust mite Dermatophagoides pteronyssinus (Trouessart): a novel gene arrangement among arthropods. Bmc Genomics 2009;10

13 Chang DD, Clayton DA: Priming of Human Mitochondrial-DNA Replication Occurs at the Light-Strand Promoter. Proceedings of the National Academy of Sciences of the United States of America 1985;82:351-355.

14 Chan TF, Ji KM, Yim AK, Liu XY, Zhou JW, Li RQ, Yang KY, Li J, Li M, Law PT, Wu YL, Cai ZL, Qin H, Bao Y, Leung RK, Ng PK, Zou J, Zhong XJ, Ran PX, Zhong NS, Liu ZG, Tsui SK: The draft genome, transcriptome, and microbiome of Dermatophagoides farinae reveal a broad spectrum of dust mite allergens. J Allergy Clin Immunol 2015;135:539-548.

15 Chaisiri K, McGarry JW, Morand S, Makepeace BL: Symbiosis in an overlooked microcosm: a systematic review of the bacterial flora of mites. Parasitology 2015;142:1152-1162.

16 Mofiz E, Holt DC, Seemann T, Currie BJ, Fischer K, Papenfuss AT: Genomic resources and draft assemblies of the human and porcine varieties of scabies mites, Sarcoptes scabiei var. hominis and var. suis. Gigascience 2016;5:23.

17 Arlian LG, Morgan MS, Rider SD: Sarcoptes scabiei: genomics to proteomics to biology. Parasite Vector 2016;9:380.

18 Lowe TM, Eddy SR: tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 1997;25:955-964.

**Supplemental Figure Legends**

**Supplemental Figure 1. Kmer analysis of *D. pteronyssinus* Illumina data**

Kmer analysis of two of the Illumina datasets used in this study. A) MiSeq, B) HiSeq2500, C) Illumina sequence from *S. scabiei* (Rider genome), D) unpublished sequence data from a haploid fungal genome. Heterozygosity is indicated as blue line is the distribution of heterozygous kmers, green line homozygous kmers

**Supplemental Figure 2. Dotplot of mitochondrial genomes**

On the horizontal axis is the sequence of the published *D. pteronyssinus* mitochondrial genome and on the vertical axis is the mitochondrial genome assembled from the PacBio sequence used in this study.

**Supplemental Figure 3. Multiple sequence alignment of the mitochondrial D loop region**

The D loop region of both the published mitochondrial genome of *D. pteronyssinus* and our assembled mitochondrial genome is highlighted.

**Supplemental Figure 4. Dotplot of *Serratia marcescens* genomes**

On the horizontal axis is the sequence from *S. marcescens* WW4 and on the vertical axis is the *S. marcescens* genome assembled from the PacBio sequence used in this study.

**Supplemental Figure 5. Predominant bacterial species in the DP microbiome**

The relative percentage of the top ten species found in the DP microbiome is shown as a pie chart.