



Chromosome-level assembly of *Dermatophagoides farinae* genome and transcriptome reveals two novel allergens Der f 37 and Der f 39

Jiajie Chen^a, Zelang Cai^a, Dingding Fan^b, Jiayu Hu^a, Yibo Hou^a, Yongsen He^a, Zhen Zhang^a, Zhenfu Zhao^a, Pan Gao^c, Wanzhen Hu^c, Jinlyu Sun^{d**}, Jiang Li^{b***} and Kunmei Ji^{a*}

ABSTRACT

Accurate house dust mite (HDM) genome and transcriptome data would promote our understanding of HDM allergens. We sought to assemble chromosome-level genome and precise transcriptome profiling of *Dermatophagoides farinae* and identify novel allergens. In this study, genetic material extracted from HDM bodies and eggs were sequenced. Short-reads from next generation sequencing (NGS) and long-reads from PacBio/Nanopore sequencing were used to construct the *D. farinae* nuclear genome, transcriptome, and mitochondrial genome. The candidate homologs were screened through aligning our assembled transcriptome data with amino acid sequences in the WHO/IUIS database. Our results showed that compared with the *D. farinae* draft genome, bacterial DNA content in the presently developed sequencing reads was dramatically reduced (from 22.9888% to 1.5585%), genome size was corrected (from 53.55 Mb to 58.77 Mb), and the contig N50 was increased (from 8.54 kb to 9365.49 kb). The assembled genome has 10 contigs with minimal microbial contamination, 33 canonical allergens and 2 novel allergens. Eight homologs ($\geq 50\%$ homology) were cloned; 2 bound HDM allergic-sera and were identified as allergens (Der f 37 and Der f 39). In conclusion, a chromosome-level genome, transcriptome and mitochondrial genome of *D. farinae* was generated to support allergen identification and development of diagnostics and immunotherapeutic vaccines.

Keywords: *Dermatophagoides farinae*, Dust mite, Chromosome-scale genome, Nanopore sequencing, Transcriptome sequencing, Der f 37, Der f 39

LETTER TO THE EDITOR,

House dust mites (HDMs), especially *Dermatophagoides pteronyssinus* (Der p) and *Dermatophagoides farinae* (Der f), are major sources of inhaled allergens.¹⁻⁴ In-depth analyses of the full spectrum of HDM allergens are needed to elucidate HDM allergy mechanisms and guide diagnostic and immunotherapeutic development.⁵

Publication of a Der f draft genome produced with next generation sequencing (NGS) facilitated the identification of numerous allergens,⁶⁻¹¹

^aDepartment of Biochemistry and Molecular Biology, Laboratory Department of Pinghu Hospital, Health Science Center, Shenzhen University, Shenzhen, 518060, China

*Corresponding author. **Corresponding author. ***Corresponding author. sunjinlv@pumch.cn or Jiang.Lee2018@gmail.com or jkm@szu.edu.cn

Full list of author information is available at the end of the article <https://doi.org/10.1016/j.waojou.2021.100590>

<http://doi.org/10.1016/j.waojou.2021.100590>

Received 25 June 2021; Received in revised form 1 September 2021; Accepted 3 September 2021

designated by the World Health Organization/International Union of Immunological Societies Allergen Nomenclature Sub-committee (WHO/IUIS).¹² However, the Der f draft genome has shortcomings due to technical limitations.⁵ For example, the Der f 23 cDNA sequence differs from its draft genome corollary.¹³ Because microbiota sequences were removed manually, the draft may contain microbiome sequences.⁵ Additionally, limitations of short-read sequencing were likely to produce much scaffold gaps.⁵

During genome assembly, it is important to minimize cross-species DNA contamination.¹⁴ Herein, we conducted DNA sequencing of HDM eggs with little microbial genetic contamination. Long-read sequencing with PacBio and Nanopore was performed to obtain a chromosome-level assembly. Homology comparison was performed to optimize transcriptome accuracy. Novel HDM allergen candidates were evaluated with specific immunoglobulin (Ig)E-binding assays.

Due to the symbiotic relationship in the digestive tract, it was not possible to obtain pure mite bodies without any microorganisms through aseptic culture methods.⁵ Our attempts to isolate Der f cells aseptically were not successful (data not shown). To minimize microbial DNA contamination, we isolated HDM eggs (Fig S1A, B) by centrifugation with a density gradient solution and extracted genomic DNA for short-read sequencing. A library was constructed from 500-base pair (bp) fragments, generating a total of 43.7 Gb of data (Table S1). Read assignment analysis showed that the bacterial content was reduced by 93% in reads obtained from eggs (1.5585%) compared to that from bodies (22.9888%). The main bacterial species contaminating body reads were *Proteobacteria* (19.24%), *Bacteroidetes* (2.31%), and *Firmicutes* (0.78%). The main species in the egg reads were also *Proteobacteria* (0.58%), *Bacteroidetes* (0.63%) and *Firmicutes* (0.18%) (Fig S1C, D). The eukaryotic read content was improved by 27% (eggs, 98.3437% vs. bodies, 76.8943%; FIG S1E, F). These data indicate that sequencing reads obtained from HDM eggs constituted a good template for genome assembly.

To reduce gaps, 25.7 Gb of raw long-read sequencing data were obtained from Nanopore

sequencing with 400-fold coverage of the estimated genome size. The N50 length of the raw Nanopore reads was 23.9 kb (Table S2). Because long reads may contain errors, assembly data (~8.3 Gb) were obtained by filtering and correcting ~22.7 Gb of raw read data (Table S2). Our assembly strategy is summarized in Fig. S2. Using read data from Der f-egg genomic DNA as a template, with an identity criterion of $\geq 90\%$, we obtained an assembly with no large gaps and an average sequence depth that had half of the genome sequence coverage as hybrid sequences. We removed 6 hybrid sequences that totaled 104,483 bp.

The assembled genome was submitted to the National Center for Biotechnology Information (BioProject ID PRJNA512594; accession no. SDOV00000000). Based on the updated assembly, the Der f genome size was corrected from 53.55 Mb to 58.77 Mb, with 10 contigs (Table S3). Contig N50 was increased from 8.54 kb to 9365.49 kb, and the contig N90 quantity was decreased from 6350 to 8 (Table S4). The final genome obtained with Nanopore sequencing consists of 10 contigs and a circular mitochondrial DNA (Fig. 1). With the exception of Contig2, all sequences exceeded 2 Mb, with the longest one exceeding 13 Mb, indicating that the assembly quality reached a near-chromosome level (Table S5).

To construct an allergen gene map (Fig. 1A), we annotated 33 canonical allergen genes in the assembled genome, including 2 newly discovered proteins, namely Der f 37 and Der f 39, in corresponding contig positions. To obtain a high-quality gene set, we performed homology, next generation RNA sequencing (RNA-seq), and *de novo*-based genome annotation of the chromosome-level assembly. For RNA-seq, we obtained 10.66 Gb of RNA-seq reads from mite bodies and 41,602 transcripts from a PacBio Iso-Seq assembly with an N50 size of 2627 (Table S6). We identified 10,684 protein-coding genes (mean exons per gene, 3.85; mean gene length, 2638 bp; and mean complete coding sequence length, 1669 bp; Table S7). More than 91.67% of the identified genes were functionally annotated via searches of the NCBI non-redundant protein, SwissProt, and KEGG databases. This reduced quantity of genes, compared

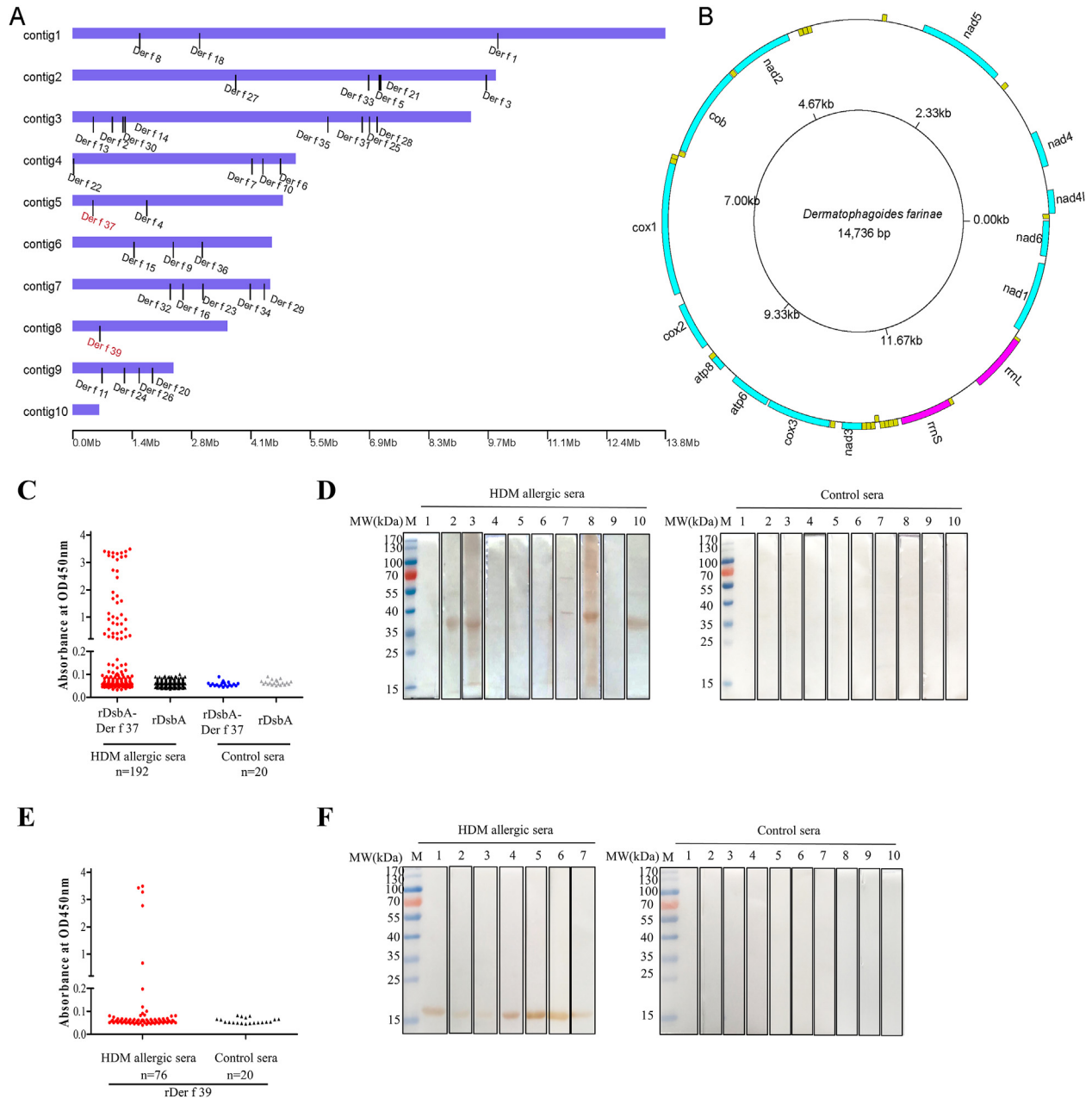


Fig. 1 Chromosome-level assembly of genome and mitochondrion genome of *D. farinae* and identification of novel HDM allergens Der f 37 and Der f 39. **A.** Genome assembly showing locations of allergen genes. **B.** Mitochondrial genome of *D. farinae*. Mite allergens based on WHO/IUIS allergen nomenclature (<http://www.allergen.org/search.php?TaxSource=Animalia%20Arthropoda>). Der f 17 sequence information: not available; HDM allergen Groups 12 and 19: not found; Der p 38 homolog in *D. farinae* (bacterial lytic enzyme like protein, GenBank accession No. MT360919.1) showed no IgE-binding activity in this study. **C.** IgE binding activity determined by IgE-western blots of rDsbA-Der f 37 with individual sera from 192 HDM-allergic patients and 20 healthy non-allergic individuals. **D.** Western blot assay identifying rDer f 37 protein binding by IgE in sera from 10 patients with HDM allergies (left) and 10 non-HDM allergic subjects (control, right). **E.** IgE binding activity determined by IgE-ELISA of rDer f 39 with individual sera from 76 HDM-allergic patients and 20 healthy non-allergic individuals. **F.** Western blot assay identifying rDer f 39 protein binding by IgE in sera from 7 patients with HDM allergies (left) and 10 non-allergic subjects (control, right). The HDM-specific IgEs (>100 kU_A/L) within the sera samples were evaluated using an ImmunoCAP system.

to the 16,145 genes in the prior draft genome, indicates that the updated assembly contains fewer contaminant and fragmented genes (Table S8). Through integration of all predicted

repeat results, about 9.7% of the genome could be attributed to transposable elements (TEs) and the highest content of family was DNA (Tables S9 and S10). We updated the

mitochondrial genome assembly, which was found to encode 37 genes, including 13 protein-coding genes, 2 rRNA genes, and 22 tRNA genes (Fig. 1B).

Assessment of the quality of our assembly and annotation in BUSCO (Benchmarking Universal Single-Copy Orthologs) indicated that our current Nanopore and prior NGS assemblies were 96.70% and 93.40% complete, respectively. Gene set completeness levels were 98.40% and 94.60% for the current Nanopore and prior NGS assemblies, respectively, indicating that the new one has 12.60% greater gene completeness than the prior draft. Moreover, the level of gene completeness of the present assembly exceeds that obtained in prior *Arachnida* genome efforts, including those for *Ixodes scapularis* (78.80%), *Stegodyphus mimosarum* (81.20%), and *Tetranychus urticae* (92.40%) (Table S11).

The dramatically improved assembly statistics obtained here relative the original draft genome is consequential because high-quality transcript data are conducive to allergen gene discovery. Our cloned Der f 23 cDNA sequence is same as the Der f 23 sequence in the assembled transcriptome, but different from that in the former draft genome.¹³ This improvement can be attributed to our combined use of multiple sequencing methods with complementary technical advantages that facilitated the rapid accurate *de novo* assembly (Fig. S2).¹⁵

As of October 2019, 959 allergens had been collected in the WHO/IUIS allergen database. We used homology analysis to align our assembled transcriptome data with amino acid sequences in the WHO/IUIS database. With a BLAST filter of identify $\geq 50\%$, 29 homologs were filtered out (Table S12). The first 8 candidate homologs of interest were cloned and expressed for identification of allergenicity (Table 1). The results showed that petrotrophic-like protein in Der f (Genbank No. MK419030.1) has 74.90% homology with Der p 37 allergen gene (Genbank No. AVD73319.1). We obtained positive IgE-binding results with specific IgEs from the serum of individuals with strong HDM allergies (slgE) for Der f petrotrophic-like protein [positive rate: 21.4%, 41/192 in IgE-enzyme-linked immunosorbent assay

(ELISA); 4/10 in IgE-western blot (WB); 4/10 in IgE-dot ELISA] (Fig. 1C-D and Fig. S3). Troponin C-like protein from Der f (Genbank No. MK419032.1) was found to have 95.42% homology with the allergen Tyr p 34 (Genbank No. ACL36923.1), and a positive slgE-binding reaction (positive rate: 9.21%, 7/76 in IgE-ELISA; positive serum: 7/7 in IgE-WB; 6/7 in IgE-dot ELISA) (Fig. 1E-F and Fig. S4). Based on these results affirming that these 2 homologs are novel Der f allergens, they have been named Der f 37 and Der f 39 by WHO/IUIS, respectively (Table 1). We did not observe slgE binding for the remaining 6 homologs, and thus can infer they are unlikely to be allergens (Table 1). Finally, we retrieved complete gene sequences and genomic location of 33 canonical HDM allergens and 2 novel HDM allergens encoded in the assembled *D. farinae* genome (Table S13).

Allergen homologs, including homologs of panallergens, can be considered potential allergen candidates.¹⁶ We obtained 6 candidates with amino acid sequence homologies to canonical allergens ranging from 50.63% to 84.83% (Table 1). These candidates are: heat shock cognate 70-like protein (GenBank No. MT360915.1), homolog of Aed a 8; bacterial lytic enzyme-like protein (GenBank No. MT360919.1), homolog of Der p 38; cytochrome c-like protein (GenBank No. MT360914.1), homolog of Cur l 3; peptidyl-prolyl *cis-trans* isomerase-like protein (GenBank No. MT360916.1), homolog of Asp f 27; lysosomal aspartic protease like protein (GenBank No. MT360917.1), homolog of Aed a 11; and aldehyde dehydrogenase-like protein (GenBank No. MT360918.1), homolog of Cla h 10. These 6 homologs exhibited no binding activity with slgE ($N = 15$; >100 kU_A/L), nor with IgEs from non-HDM allergic individuals ($N = 10$) (Figs. S5, S6). To further confirm whether the Der p 38 homolog bacterial lytic enzyme-like protein has slgE binding activity, an additional slgE binding assay was conducted with an expanded sample of HDM allergic sera ($N = 100$). Similarly, an *Escherichia coli*-derived recombinant protein of the Der p 38 homolog showed no slgE-binding activity ($N = 100$) (Fig. S6). Der p 38 (GenBank No. MT360919.1) differs from Der f 38 (GenBank No. QHQ72282.1) by 2 amino acids (Fig. S7). Both

Locus tag (GenBank No.)	Biochemical function	Deduced no. amino acids	Homolog (GenBank No.) % similarity species	Allergenicity	WHO/IUIS Allergen Nomenclature ^b
Dfarinae12240 (MK419032.1)	Troponin C-like protein	153	Tyr p 34 (ACL36923.1) 95.42% <i>Tyrophagus putrescentiae</i>	IgE-ELISA 9.21% (7/76); IgE-WB ^a 100% (7/7); IgE- dot blot ^a 85.71% (6/7)	Der f 39
Dfarinae12320 (MT360915.1)	Heat shock cognate 70-like protein	202	Aed a 8 (ABF18258.1) 84.83% <i>Aedes aegypti</i>	No activity	None
Dfarinae07901 (MT360919.1)	Bacterial lytic enzyme-like protein	150	Der p 38 (AAN02509.1) 80.00% <i>Dermatophagoides pteronyssinus</i>	No activity	None
Dfarinae06505 (MK419030.1)	Petrotrophic-like protein	250	Der p 37 (AVD73319.1) 74.90% <i>Dermatophagoides pteronyssinus</i>	IgE-ELISA 21.4% (41/192); IgE- WB 40% (4/10); IgE-dot blot 40% (4/10)	Der f 37
Dfarinae09175 (MT360914.1)	Cytochrome c- like protein	106	Cur l 3 (AAK67492.1) 66.99% <i>Curvularia lunata</i>	No activity	None
Dfarinae11953 (MT360916.1)	Peptidyl-prolyl cis-trans isomerase-like protein	227	Asp f 27 (CAI78448.1) 59.15% <i>Aspergillus fumigatus</i>	No activity	None
Dfarinae11869 (MT360917.1)	Lysosomal aspartic protease-like protein	401	Aed a 11 (XP_001657556.1) 54.36% <i>Aedes aegypti</i>	No activity	None
Dfarinae12442 (MT360918.1)	Aldehyde dehydrogenase- like protein	490	Cla h 10 (CAA55072.2) 50.63% <i>Cladosporium herbarum</i>	No activity	None

Table 1. Identification of HDM-specific IgE binding activity of 8 allergen homologs in the *D. farinae* genome. ^aHDM allergic sera from positive samples in IgE-ELISA assay. ^bWHO/IUIS database at <http://www.allergen.org/>, not available; BLAST, Basic Local Alignment Search Tool, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

have yielded positive skin prick test results and thus were recognized as allergens by WHO/IUIS despite there being no published IgE-binding assays (Der p 38: <http://www.allergen.org/viewallergen.php?aid=949> and Der f 38: <http://www.allergen.org/viewallergen.php?aid=1014>).

The remaining 21 additional proteins with $\geq 50\%$ homologous amino acid sequence require further allergenicity probing (Table S12).

In summary, we used multiple sequencing technologies to assemble a Der f chromosome-level genome and transcriptome. We identified 2

previously unknown allergens: Der f 37 and Der f 39. These findings will be helpful for clarifying pathogenic mechanisms of HDM allergies and for supporting the development of diagnostic methods and immunotherapeutic vaccines.

Abbreviations

House dust mites (HDMs); *Dermatophagoides pteronyssinus* (Der p); *Dermatophagoides farinae* (Der f); Next generation sequencing (NGS); World Health Organization/International Union of Immunological Societies Allergen Nomenclature Sub-committee (WHO/IUIS).

Funding

The present study was supported in part by research funding from the National Natural Science Foundation of China (Grants No. 81571570, 82071806, 30671943 and 30671943), SZU Top Ranking Project (Grant No. 86000000210) and Shenzhen City (Grants No. JSGG20200102165803939 and JSGG20200225151806035).

Authors' consent for publication

All the authors consent the publication of the manuscript.

Author contributions

JC, ZC, JH, YBH, YSH and ZZ performed experiments and interpreted results. JC, YH, ZZ, PG, WZ, DF and JL contributed to the data analysis. JC, JS, and KJ supervised the projects and participated in experimental design and technical discussions. JC, DF and ZZ wrote the paper. ZFZ, JS, JL and KJ revised the manuscript.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Permission to conduct this study was obtained from the Ethics Committee of the First Affiliated Hospital of Guangzhou Medical College (No. 2012-51). Informed consent was obtained from all individual participants included in the study. All procedures involving human participants were in accordance with the ethical standards of the committee.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

We thank Miss Li-Zhu Yang from Ji Kunmei lab for her technologic support in the protein expression and purification and other members of Ji Kunmei lab for their critical comments in the manuscript preparation.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.waojou.2021.100590>.

Author details

^aDepartment of Biochemistry and Molecular Biology, Laboratory Department of Pinghu Hospital, Health Science Center, Shenzhen University, Shenzhen, 518060, China
^bEasyATCG L.L.C, Shenzhen, 518000, China ^cShenzhen University General Hospital, Shenzhen, 518060, China ^dAllergy Department, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, 100730, China.

REFERENCES

1. Pawankar R, Canonica GW, Holgate ST, Lockey RF, Organization WA. *World Allergy Organization (WAO) White Book on Allergy*. Milwaukee (WI): World Allergy Organization; 2011.
2. Voorhorst R, Spieksma FTM, Varekamp H, Leupen MJ, Lyklema AW. The house-dust mite (*Dermatophagoides pteronyssinus*) and the allergens it produces. Identity with the house-dust allergen. *J Allergy*. 1967;39:325-339.
3. Miyamoto T, Oshima S, Ishizaki T, Sato SH. Allergenic identity between the common floor mite (*Dermatophagoides farinae* Hughes, 1961) and house dust as a causative antigen in bronchial asthma. *J Allergy*. 1968;42:14-28.
4. Thomas WR, Smith WA, Hales BJ, Mills KL, O'Brien RM. Characterization and immunobiology of house dust mite allergens. *Int Arch Allergy Immunol*. 2002; 129:1-18.
5. Chan TF, Ji KM, Yim AK, et al. The draft genome, transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens. *J Allergy Clin Immunol*. 2015;135(2):539-548.
6. Lin J, Li M, Liu Y, et al. Expression, purification and characterization of Der f 27, a new allergen from *Dermatophagoides farinae*. *Am J Transl Res*. 2015;7(7):1260-1270.
7. An S, Chen L, Long C, et al. *Dermatophagoides farinae* allergens diversity identification by proteomics. *Mol Cell Proteomics*. 2013;12(7):1818-1828.
8. Li Y, Wang Y, Ran P, Yang P, Liu Z. IgE binding activities and in silico epitope prediction of Der f 32 in *Dermatophagoides farinae*. *Immunol Lett*. 2019;213:46-54.
9. ElRamlawy KG, Fujimura T, Baba K, Kim JW, Kawamoto C, Isobe T, et al. Der f 34, a Novel Major House Dust Mite Allergen Belonging to a Highly Conserved Rid/YjgF/YER057c/UK114 Family of Imine Deaminases. *J Biol Chem*. 2016;291(41): 21607-21615.
10. Fujimura T, Aki T, Isobe T, et al. Der f 35: an MD-2-like house dust mite allergen that cross-reacts with Der f 2 and Pso o 2. *Allergy*. 2017;72(11):1728-1736.
11. Bordas-Le Floch V, Le Mignon M, Bussi eres L, et al. A combined transcriptome and proteome analysis extends the allergome of house dust mite *Dermatophagoides* species. *PLoS One*. 2017;12(10), e0185830.
12. Pom es A, Davies JM, Gadermaier G, et al. WHO IUIS allergen nomenclature sub-committee. WHO/IUIS allergen

- nomenclature: providing a common language. *Mol Immunol*. 2018;100:3-13.
13. He Y, Dou C, Su Y, et al. Identification of Der f 23 as a new major allergen of *Dermatophagoides farinae*. *Mol Med Rep*. 2019;20(2):1270-1278.
 14. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2014;2: e675.
 15. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*. 2016;32(7):1009-1015.
 16. Chruszcz M, Kapingidza AB, Dolamore C, Kowal K. A robust method for the estimation and visualization of IgE cross-reactivity likelihood between allergens belonging to the same protein family. *PLoS One*. 2018;13(11), e0208276.

1 **Supplementary Informations**

2 Chromosome-level assembly of *Dermatophagoides farinae* genome and
3 transcriptome reveals two novel allergens Der f 37 and Der f 39

4

5 Table of Contents:

6 1. Supplementary materials and methods

7 2. Supplementary figure legends

8 FIGURE S1.

9 FIGURE S2.

10 FIGURE S3.

11 FIGURE S4.

12 FIGURE S5.

13 FIGURE S6.

14 FIGURE S7.

15 3. Supplementary table legends

16 TABLE S1.

17 TABLE S2.

18 TABLE S3.

19 TABLE S4.

20 TABLE S5.

21 TABLE S6.

22 TABLE S7.

23 TABLE S8.

24 TABLE S9.

25 TABLE S10.

26 TABLE S11.

27 TABLE S12.

28 TABLE S13.

1. Supplementary materials and methods

Mite culture and egg isolation

D. farinae culture and purity check were performed as in our previous study ^[E1]. The purified mite culture mixture including bodies and eggs was dissolved with 1× phosphate buffered saline and then combined 1:1 with a density gradient solution (Histopaque®-1119, Sigma) and centrifuged at 4000 rpm for 5 min. After centrifugation, the solution was divided into three layers, with eggs in the middle liquid layer. The collected and purified eggs were subjected to DNA genome extraction.

Genomic DNA sample preparation and sequencing

Genomic DNA was extracted from HDM whole bodies or eggs and submitted to quality checks as described previously ^[E1]. For mite-egg genomic-DNA sequencing, NGS was performed on an Illumina HiSeq Xten platform and the library was prepared with 500-bp DNA fragment insertions. In total, 6.55 G bases were generated with pair-end 150 bp. For mite whole body genomic DNA, long-read sequencing was performed on a Nanopore Grid ION X5 platform (Oxford Nanopore Technologies) for two flow cells (R9.4) with 1D ligation library preparation techniques; a total of 25.7 Gb of raw data were collected.

RNA sample preparation and sequencing

RNA was extracted from HDM whole bodies and submitted to quality checks as in our previous study ^[E1]. *D. farinae* cDNAs were sequenced with Illumina HiSeqXten and PacBio Iso-Seq for transcriptome analysis. A total of 10.66 Gb of paired-end sequencing data was produced in Illumina HiSeqXten (insert size, ~300 bp), and 9.22 Gb of data was generated from PacBio Iso-Seq three libraries (insert sizes: 1 kbp, 2–3 kbp, and 3–6 kbp, respectively).

Microbial content proportion analysis between body and egg sequence reads

To analyze the microbial content proportion in DNA sequence reads, we selected 1

million reads randomly for mite bodies and eggs. Those reads were aligned to nucleotide of GenBank (2018.1) in BLAST2 (blastn -a 40 -e 1e-4 -m 7) ^[E2]. We assigned read origin based on the best hit with the highest alignment score. Using accession numbers of assigned reads, we converted NCBI taxonomy dump into lineages in Ncbi tax2lin (<https://github.com/zyxue/ncbitax2lin>). Read assignments were classified as *prokaryote* or *eukaryota*.

***D. farinae* genome assembly**

Raw reads with adaptor sequences were trimmed in Porechop v0.2.2 (<https://github.com/rrwick/Porechop>). Reads that were < 10 kbp with average phred score < 8 were removed by NanoFilt v2.0.02 (<https://github.com/wdecoster/nanofilt>). Filtered reads (22.7 Gb) were inputted into Canu v.1.7.03 for self-correcting, yielding ~8.3Gb of trimmed data ^[E3]. Canu v.1.7.03 parameters were set as follows: min read length = 10,000; corOutCoverage = 150; and genome size = 60 m. We select 13.1-fold coverage with length > 50 kbp. The wtdbg v1.2.8 assembler was set as follows: -edge-min 2 --rescue-low-cov-edges -p 19 -S 2 --aln-noskip to obtain an assembly.

To remove contamination and redundancy, the initial genome assembly sequences were first aligned by blastn² to nucleotide databases of Genbank (E-value $\leq 1e^{-5}$). Then, we checked carefully whether the genome sequences aligned to bacterial sequences or not ^[E2]. We aligned the NGS sequences with a sequence depth < 10, calculated by egg sequence because eggs have less bacterial contamination, to protein databases of Genbank using blastx (E-value $\leq 1e^{-3}$) to examine which species matched ^[E2]. All matched species were eukaryote, proving that there was no contamination. At this step, we removed nine polluted sequences leaving 1,734,897 bp in total. Sequences shorter than 1 Mbp were aligned to the whole primary assembly in blat v. 35 ^[E4] with an identity criterion of $\geq 90\%$. There were no big gaps. Sequence average depth was half of the genome sequence coverage as hybrid sequences. At this step, we removed six hybrid sequences, leaving 104,483 bp.

Finally, the clean Nanopore reads longer than 50 kb were used to construct a scaffold in SSPACE-LongRead v1-1 with the following parameters: -i 85 -l 5 -o 5000

[E5]. Pbjelly v15.8.24 was used to fill gaps and further splice with the following parameters: -minMatch 8 -minPctIdentity 85 -bestn 1 -nCandidates 20 -maxScore -500 -nproc 32 -noSplitSubreads [E6]. Additionally, LR_Gapcloser was employed to process gap close and result in no scaffold gap [E7]. The longest 100X Nanopore clean reads were aligned to the scaffold with minimap2 v2.12-r836-dirty, and racon v1.3.1 (<https://github.com/isovic/racon>) was used to polish the scaffold five times iteratively [E8]. Illumina NGS clean reads from eggs were aligned to the scaffold with bwa v 0.7.17-r1188 [E9]. After removal of duplicated reads from polymerase chain reactions with Picard 1.114 (PICARD/MarkDuplicates.jar, <https://broadinstitute.github.io/picard/>), the bwa alignment results were re-aligned with GATK 3.4-46 IndelRealigner (<https://gatk.broadinstitute.org/hc/en-us>). Lastly, the scaffolds were polished with pilon v1.22 (<https://github.com/broadinstitute/pilon>) three times iteratively to obtain the final assembly [E10].

***D. farinae* genome annotation**

We performed repetitive element annotation. Tandem repeats were predicted using Tandem Repeats Finder v4.04 (<http://tandem.bu.edu/trf/trf.html>). Transposable elements (TEs) were identified by homology-based and *de novo* approaches. For the homology-based approach, we first identified known TEs using Repeat Masker v4.0.7 program against Repbase v21.12 databases [E11]. We then used RepeatProtein Mask (<http://www.repeatmasker.org>) to identify TEs by aligning the genome sequence to the TE protein database [E11]. For the *de novo* method, we constructed a repeat library generated by RepeatScout v1.0.5 [E12] and ltr_finder v1.07 [E13], then RepeatMasker was used to identify the repeat sequences. In total, 9.7% of the genome bases could be attributed to TEs; among them, the highest content family was DNA transposon.

Then, the soft-mask genome assembly based on transposable elements were further used for gene prediction. The final gene prediction was integrated from three methodologies: *ab initio*, homology, and transcriptome.

(1) *Ab initio* gene prediction

Ab initio prediction was performed based on the soft-masked genome from transposons.

Four programs (AUGUSTUS v3.2.3, snap v2006-07-28, GeneMark v4.35, and fgenesh from MolQuest package v 2.4.5) were used for prediction.

(2) Gene prediction based on homolog evidence

The protein sequences (<ftp://ftp.ensemblgenomes.org/pub/metazoa/release-46>) of five related representative species (*Tetranychus urticae*, *Drosophila melanogaster*, *Daphnia pulex*, *Ixodes scapularis*, and *Stegodyphus mimosarum*) and published protein sequences of *D. farinae* were aligned to the assembly using BLAST v2.2.26 (-p tblastn -e 1e-05) ^[E1]. Alignment and identification of accurately spliced alignments was accomplished in Gene Wise v2.2.0.

(3) Gene prediction from transcriptome evidence

The cDNAs that were generated by Pacbio sequencing were aligned against the assembly genome in PASA (identity ≥ 0.95 , coverage ≥ 0.95) (<https://pasapipeline.github.io/>) to predict the possible gene model. Transcript clean reads generated from Illumina sequence platform were aligned to the improved genome assembly in HISTAT v2.1.0, and transcript assembly was conducted with StringTie v1.3.2.

(4) Final gene sets from three sources of evidence

Gene sets obtained as described above were integrated with Evidence Modeler v1.1.1. For genes supported by only the *ab initio* method, we retained putative genes that had at least one protein domain with an InterPro annotation or that had RNA-seq reads mapped with a fragments per kb per million fragments (FPKM) greater or equal to that calculated by StringTie v1.3.2. Finally, a total of 12,872 genes were obtained. Gene functions were assigned based on best-matched hits to NCBI non-redundant protein (NR), SwissProt, and Kyoto Encyclopedia of Genes and Genomes (KEGG) with BLAST (-p blastp -e 1e-5) ^[E2]. Gene motifs and domains were identified by InterProScan v 5.27 referencing protein databases. The overlap length of each gene's coding region was calculated based on predicted genes from the above *ab initio*, homology, and transcript methods. All genes having $\geq 50\%$ overlap by at least one method were retained. According to a Venn diagram analysis of evidence, most genes were derived from homology evidence and transcriptome evidence.

Mitochondrion assembly and annotation

First, blasr v5.1 was used to align the clean Nanopore reads to the *D. farinae* mitochondrial sequence (NC_013184.1) assembled by NGS technology with the following parameters: --minMatch 8 --bestn 10 --noSplitSubreads --advanceExactMatches 1 --nCandidates 1 --maxAnchorsPerPosition 1 --sdpTupleSize 7. Next, the reads which alignment length ≥ 1 kb and identity $\geq 90\%$ were selected for Canu v.1.7.0^[E3] assembly with the parameters set as follows: minReadLength = 1000 minOverlapLength=500; corMhapSensitivity = normal; corMinCoverage = 4; corOutCoverage = 40; correctedErrorRate = 0.144; corMaxEvidenceCoverageGlobal = 1.0 $\times$; corMaxEvidenceCoverageLocal = 2.0 $\times$; corPartitions = 30; ovsMethod = sequential trimReadsCoverage = 1. Finally, the initial assembly generated by canu was polished by pilon v 1.22 three times iteratively to obtain a 14,736 bp assembly. Protein-coding genes and rRNAs were identified with reference to MITOS (<http://mitos.bioinf.uni-leipzig.de/index.py>) using the invertebrate genetic code. tRNAs were identified by sequence similarity to *D. farinae* mitochondrion genes (NC_013184.1) by BLAST (-p blastn -e 1e-10).

Analysis of genome assembly completeness

Nanopore clean reads ≥ 50 kb were aligned to the assembly with Graphmap v 0.5.2 and visualized with Tablet. We checked read coverage of the assembly manually^[E14], and found that all reads aligned properly and continuously, affirming the accuracy of assembly. BUSCO (version 3) has been widely used for the assessment of genome assemblies and gene sets based on evolutionarily informed expectations of gene content^[E15]. To evaluate the completeness of the *D. farinae* genome, BUSCO v 3.0.2 was used in the genome mode and protein mode with eukaryota_odb9 ortholog gene sets.

Homologous alignment of amino acid sequences with allergens named by WHO/IUIS

As of October 2019, the allergen online database of the WHO/IUIS has collected 959

allergens (<http://www.allergen.org/>). To identify the allergen genes on our assembled HDM genome, BLAST (-p tblastn -e 1e-3) in-house and alignment coverage were calculated^[E3]. We chose homologous genes with amino acid sequences identity $\geq 50\%$ as potential allergen homologs.

Expression of eight potential allergen homolog genes, protein purification, and IgE-binding assay

We found eight potential allergen homolog cDNAs (encoding troponin C-like protein, heat shock cognate 70-like protein, bacterial lytic enzyme-like protein, petrotrophic-like protein, cytochrome c-like protein, peptidyl-prolyl cis-trans isomerase-like protein, lysosomal aspartic protease-like protein, and aldehyde dehydrogenase-like protein) in *D. farinae* that were present in HDM cDNA libraries; their cDNA sequences matched the results of our transcriptome data, verified by Sanger DNA sequencing. The corresponding amino acid sequences were reported in GenBank (Accession nos. MK419032, MT360915, MT360919, MK419030, MT360914, MT360916, MT360917, and MT360918).

These eight recombinant proteins were expressed and purified according to our protocol^[E1, E16, E17]. The recombinant expression plasmids of troponin C-like protein, heat shock cognate 70-like protein, petrotrophic-like protein, lysosomal aspartic protease-like protein, and aldehyde dehydrogenase-like protein were constructed with pET-His vectors. The recombinant expression plasmids of bacterial lytic enzyme-like protein, cytochrome c-like protein, and peptidyl-prolyl cis-trans isomerase like-protein were constructed with pET-DsbA vectors. The purified recombinant proteins were subjected IgE-binding assays (western blots, dot blots, and ELISAs) according to our protocol^[E1, E16, E17]. All procedures involving human participants were conducted in accordance with ethical standards. Ethics approval was obtained from the First Affiliated Hospital of Guangzhou Medical College. For statistical analysis, non-normally distributed data were compared between groups with Wilcoxon's rank-sum test (significance at $p < 0.05$). SAS 9.1 for Windows software (SAS Institute, Cary, North Carolina, USA) was used for all analyses.

209 **References**

- 210 E1. Chan TF, Ji KM, Yim AK, Liu XY, Zhou JW, Li RQ, et al. The draft genome,
211 transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum
212 of dust mite allergens. *J Allergy Clin Immunol* 2015; 135:539-48.
- 213 E2. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc*
214 2007; 2007:pdb. top17.
- 215 E3. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu:
216 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat
217 separation. *Genome Res.* 2017; 27(5): 722-736.
- 218 E4. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002; 12(4): 656-
219 64.
- 220 E5. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes
221 using long read sequence information. *BMC Bioinformatics.* 2014; 15:211.
- 222 E6. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid
223 JG, Worley KC, Gibbs RA. Mind the gap: upgrading genomes with Pacific Biosciences
224 RS long-read sequencing technology. *PLoS One.* 2012; 7(11): e47768.
- 225 E7. Xu GC, Xu TJ, Zhu R, Zhang Y, Li SQ, Wang HW, Li JT. LR_Gapcloser: a tiling
226 path-based gap closer that uses long reads to complete genome assembly. *Gigascience.*
227 2019; 8(1):giy157.
- 228 E8. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.*
229 2018; 34(18):3094-3100.
- 230 E9. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler
231 transform. *Bioinformatics.* 2010; 26(5):589-95.
- 232 E10. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al.
233 Pilon: an integrated tool for comprehensive microbial variant detection and genome
234 assembly improvement. *PLoS One.* 2014; 9(11): e112963.
- 235 E11. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive
236 elements in eukaryotic genomes. *Mob DNA.* 2015; 6:11.
- 237 E12. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in
238 large genomes. *Bioinformatics.* 2005; 21 Suppl 1:i351-8.
- 239 E13. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-
240 length LTR retrotransposons. *Nucleic Acids Res.* 2007; 35(Web Server issue): W265-
241 8.
- 242 E14. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, et al. Using
243 Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform.*
244 2013; 14(2):193-202
- 245 E15. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov
246 G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and
247 Phylogenomics. *Mol Biol Evol.* 2018; 35(3):543-548.
- 248 E16. Zhang Z, Cai Z, Hou Y, Hu J, He Y, Chen J, et al. Enhanced sensitivity of capture
249 IgE-ELISA based on a recombinant Der f 1/2 fusion protein for the detection of IgE
250 antibodies targeting house dust mite allergens. *Mol Med Rep.* 2019 May;19(5):3497-

251 3504.
252 E17. He Y, Dou C, Su Y, Chen J, Zhang Z, Zhao Z, et al. Identification of Der f 23 as
253 a new major allergen of *Dermatophagoides farinae*. Mol Med Rep. 2019; 20(2):1270-
254 1278.
255
256

2. Supplementary figure legends

FIGURE S1. Microbial content proportion of *D. farinae* egg genome. Microscope observation of egg with low (100×; **A**) and high (400×; **B**) magnification. Comparison of prokaryotic read content (**C**) and species (**D**) between egg- and body-derived genome sequence reads. Comparison of eukaryotic reads content (**E**) and species (**F**) between dust mite eggs and bodies genome sequence reads.

FIGURE S2. Pipelines of hybrid assembly of short and long reads for *D. farinae* genome assembly, annotation and evaluation. **A.** Flow chart of genome assembly and annotation of *D. farinae*. **B.** Flow chart of *D. farinae* genome evaluation.

FIGURE S3. Identification of Der f 37 as a novel HDM allergen. **A.** Deduced amino acid sequence of petrotrophic-like protein (GenBank accession no. MK419030). Synthesized Der f 37 cDNA was subcloned into pET expression system to generate Der f 37 (theoretical molecular mass: ~36 kDa). **B.** Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) of purified rDer f 37 from (a) *E. coli* BL21 cells with lambda CE6 infection or (b) BL21(DE3) cells stained with Coomassie brilliant blue. **C.** IgE binding activity determined by IgE-western blots of rDsbA-Der f37 ($Z = 5.198$, $p < 0.01$) with individual sera from 192 HDM-allergic patients and 20 healthy non-allergic individuals. Western blot (**D**) and dot blot assay (**E**) identifying rDer f 37 protein binding by IgE in sera from 10/10 patients with HDM allergies and 0/10 non-HDM allergic subjects (control).

FIGURE S4. Identification of Der f 39 as a novel HDM allergen. **A.** Deduced amino acid sequence of troponin C like protein (GenBank accession no. MK419032). Synthesized Der f 39 cDNA was subcloned into pET expression system to generate Der f 39 (theoretical molecular mass: ~17 kDa). **B.** SDS-PAGE analysis of purified rDer f39 from *E. coli* BL21(DE3) cells transformed with pET-His-Der f39 plasmid stained with Coomassie brilliant blue. **C.** IgE binding activity determined by IgE-ELISA of rDer f39 ($Z = 2.587$, $p < 0.01$) with individual sera from 76 HDM-allergic patients and 20 healthy non-allergic individuals. Western blot (**D**) and dot blot assay (**E**) identifying rDer f 39 protein binding by IgE in sera from 10/10 patients with HDM allergies and

0/10 non-allergic subjects (control). Der f 1/2 fusion protein was used as a positive control.

FIGURE S5. Recombinant expression of six candidate allergen homologs in *D. farina*.

A. Cytochrome c like protein, GenBank accession no. MT360914. **B.** Aldehyde dehydrogenase like protein, GenBank accession no. MT360918. **C.** Peptidyl-prolyl cis-trans isomerase like protein, GenBank accession no. MT360916. **D.** Lysosomal aspartic protease like protein, GenBank accession no. MT360917. **E.** Bacterial lytic enzyme like protein, GenBank accession no. MT360919. **F.** Heat shock cognate 70 like protein, GenBank accession no. MT360915) was obtained with the use of a pET expression system.

FIGURE S6. IgE binding activity of six candidate allergen homolog recombinant proteins determined by IgE-ELISA with sera from 15 HDM-allergic individuals (all, >100 kU_A/L) and 10 healthy non-HDM allergic individuals. The Der p 38 allergen homolog recombinant bacterial lytic enzyme like protein did not react with 100 individual HDM-allergic sera.

FIGURE S7.

Amino acid sequence differences between *D. farinae* bacterial lytic enzyme like protein; GenBank no. MT360919, Der f 38) found in this study and its Der p 38 homolog (GenBank no. QHQ72282.1; <http://www.allergen.org/viewallergen.php?aid=1014>).

306 **3. Supplementary table legends**

307 **TABLE S1.** Summary of *D. farinae* genome and transcriptome sequencing dataset
308 characteristics. *DNA sequence downloaded from NCBI (BioProject ID: XXX).

309 **TABLE S2.** Summary of Nanopore-sequencing statistical data. Raw read
310 means produced in Nanopore platform. Clean read means had adaptors trimmed and
311 had short and low-quality reads removed with Porechop and NanoFilt. Corrected
312 mean trimmed reads were produced in Canu ‘correct’ and ‘trim’ mode to obtain high-
313 quality reads for genome assembly.

314 **TABLE S3.** Assembled genome characteristics.

315 **TABLE S4.** Contig sequence lengths.

316 **TABLE S5.** Comparison of assembly quality between NGS/Hiseq and Nanopore
317 sequence platforms.

318 **TABLE S6.** Transcript consensus with isoform sequences determined in Pacbio
319 platform.

320 **TABLE S7.** Summary of *D. farinae* gene set characteristics.

321 **TABLE S8.** Homology and functional classification of genes annotated in *D. farinae*
322 genome.

323 **TABLE S9.** Transposable element (TE) predictions in the *D. farinae* genome.

324 **TABLE S10.** Statistics for *D. farinae* genome repeats.

325 **TABLE S11.** Assessment of genome assembly completeness in BUSCO.

326 **TABLE S12.** Allergen homologs found in the *D. farinae* genome found by searching
327 allergens from other species in the Allergen Online database ($\geq 50\%$ amino acid
328 sequence homology)

329 **TABLE S13.** Canonical allergens encoded in the assembled *D. farinae* genome.

330

331

TABLE S1.

Molecule	Platform	Source	Libraries	Bases (Gb)	Genome coverage
DNA	NGS/Hiseq Xten	Egg	1	6.55	115.30
	*NGS/Hiseq Xten	Whole body	1	9.70	170.75
	Nanopore	Whole body	1	25.76	453.47
RNA (cDNA)	NGS/Hiseq Xten	Whole body	1	10.66	187.65
	Pacbio Sequel I	Whole body	3	9.22	162.31

TABLE S2.

Characteristic	Raw	Clean	Corrected and trimmed
Total length (bp)	25,759,950,168	22,699,378,819	8,329,616,326
Genome_coveage	396	349	128
Max_length	134,855	134,803	120,414
>2 kb	1,481,407	984,435	246,466
>5 kb	1,265,176	984,435	246,466
>10 kb	986,475	984,435	246,466
Nx	Length	Length	Length
N90	9,188	13,547	25,786
N50	23,944	26,034	34,426

TABLE S3

Nuclear genome	Value
Assembly length	58.77 Mb
Number of contigs	10
Contig N50	9.27 Mb
Longest Contigs	13.80 Mb
%GC	30.4
Repeat region % of assembly	9.73
Predicted gene models	10,684
Average coding sequence length	1,669.11 bp
Average exons per gene	3.85
Mitochondrial genome	
Assembly length	14,736 bp
Protein coding gene number	13
rRNA gene number	2
tRNA gene number	22

TABLE S4.

ID	Length	GC%
ctg1	13,798,924	30.11
ctg2	9,845,874	30.28
ctg3	9,265,486	30.28
ctg4	5,189,645	30.54
ctg5	4,892,402	30.48
ctg6	4,631,692	30.92
ctg7	4,592,003	30.57
ctg8	3,602,211	30.75
ctg9	2,344,723	29.88
ctg10	613,882	31.90

TABLE S5.

Feature	NGS/Hiseq				Nanopore	
	scaffold		contig		Contig	
	Length (bp)	n	Length (bp)	n	Length (bp)	n
Max length	778,042		107,957		13,798,924	
N10	481,692	9	25,862	134	13,798,924	1
N20	368,376	22	18,791	373	13,798,924	1
N30	287,549	39	14,148	691	9,845,874	2
N40	235,409	59	11,121	1,104	9,845,874	2
N50	186,342	85	8,538	1,637	9,265,486	3
N60	143,844	118	6,673	2,324	5,189,645	4
N70	110,403	161	4,956	3,220	4,892,402	5
N80	81,164	217	3,499	4,456	4,631,692	6
N90	44,152	302	2,071	6,350	3,602,211	8
Total length	53,545,338		51,608,045		58,776,842	
Number $\geq$ 100 bp	515		11,600		10	
number $\geq$ 2000 bp	515		6,481		10	
GC%	29.40		30.50		30.40	

Hiseq platform sequences from NCBI accession no. ASGP00000000.

TABLE S6.

Max length	Total number	Average length	Number > 2000 bp	Total length	N50
8,718	41,602	2,387	25,374	99,320,767	2,627

TABLE S7.

Source	Number	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
Nanopore	10,684	1,669.11	3.85	433.29	339.89
Hiseq	16,145	1,076.38	2.60	414.57	822.30

Nanopore gene set from this project. Hiseq gene set from NCBI accession no. XXX.

TABLE S8.

Gene status		No. genes	Percentage in gene set
Total		10,684	
Annotated	InterPro	8,002	74.90%
	Swissprot	7,703	72.10%
	KEGG	7,460	69.82%
	GO	6,119	57.27%
	NR	9,671	90.52%
Unannotated		890	8.33%

NR, RefSeq non-redundant proteins (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA>).

TABLE S9.

Method	Repeat size	% of genome
Trf	1,274,524	2.17
Repeatmasker	3,566,475	6.07
Proteinmask	297,587	0.51
<i>De novo</i>	1,922,908	3.27
Total	5,718,447	9.73

TABLE S10.

Type	Repbased TEs		TE proteins		<i>De novo</i>		Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	2,760,663	4.70	10,805	0.02	39,000	0.07	2,798,041	4.76
LINE	688,312	1.17	66,470	0.11	128,157	0.22	854,680	1.45
SINE	14,422	0.02	0	0.00	0	0.00	14,422	0.02
LTR	817,898	1.39	220,312	0.37	644,566	1.10	1,421,472	2.42
Other	611	0.00	0	0.00	0	0.00	611	0.00
Unknown	0	0.00	0	0.00	1,141,531	1.94	1,141,531	1.94
Total	3,566,475	6.07	297,587	0.51	1,909,345	3.25	5,077,479	8.64

Repbased TEs: RepeatMasker analysis results using Repbase. TE proteins: RepeatProteinMask analysis results using Repbase. *De novo*: RepeatMasker analysis results using library predicted by the *de novo* method. Combined: combined results for Repbase TEs, TE proteins, and the *de novo* method.

TABLE S11.

Type	Species	Platform	BUSCO		
			Complete	Fragmented	Missing
Gene	<i>Ixodes scapularis</i>	NGS/Hiseq	78.80%	11.90%	9.30%
	<i>Stegodyphus mimosarum</i>	NGS/Hiseq	81.20%	14.90%	3.90%
	<i>Tetranychus urticae</i>	NGS/Hiseq	92.40%	2.30%	5.30%
	<i>Dermatophagoides farinae</i>	Nanopore	97.40%	1.00%	1.60%
	<i>Dermatophagoides farinae</i>	NGS/Hiseq	84.80%	9.90%	5.30%
Genome	<i>Dermatophagoides farinae</i>	Nanopore	96.00%	0.70%	3.30%
	<i>Dermatophagoides farinae</i>	NGS/Hiseq	92.10%	1.30%	6.60%

I. scapularis, *S. mimosarum*, and *T. urticae* sequences were downloaded from the Ensembl metazoa database. Hiseq platform
D. farina sequence was downloaded from NCBI (accession no. XXX).

TABLE S12.

No	Locus tag	Biochemical function	Deduced no. amino acids	Homolog allergen (GenBank no.) % similarity species	Notes
1	Dfarinae12240	Troponin C	153	Tyr p 34 (ACL36923) 95.42% <i>Tyrophagus putrescentiae</i>	Der f 39 (named in this study)
2	Dfarinae12320	Heat shock cognate 70 protein	202	Aed a 8 (ABF18258.1, 84.83%) <i>Aedes aegypti</i>	No IgE-ELISA binding
3	Dfarinae07670	Profilin	130	Tyr p 36 (AOD75399.1) 83.21% <i>Tyrophagus putrescentiae</i>	None tested
4	Dfarinae07901	Bacterial lytic enzyme protein	150	Der p 38 (AAN02509.1) 80.00% <i>Dermatophagoides pteronyssinus</i>	No IgE-ELISA binding
5	Dfarinae06505	Chitin binding protein	250	Der p 37 (AVD73319) 74.90% <i>Dermatophagoides pteronyssinus</i>	Der f 37 named in this study
6	Dfarinae10302	Beta-Enolase	433	Sal s 2 (ACH70931) 71.3% <i>Salmo salar</i>	None tested
7	Dfarinae01101	L3 ribosomal protein	398	Asp f 23 (AAM43909) 67.77% <i>Aspergillus fumigatus</i>	None tested
8	Dfarinae03903	Glucose 6-phosphate isomerase	585	Pan h 11 (AAN02509.1) 67.58% <i>Pangasianodon hypophthalmus</i>	None tested

9	Dfarinae10944	Glyceraldehyde-3-phosphate-dehydrogenase	332	Tri a 34 (CAZ76054) 67.47% <i>Triticum aestivum</i>	None tested
10	Dfarinae09175	Cytochrome c like protein	106	Cur l 3 (Q96VP3) 66.99% <i>Curvularia lunata</i>	No IgE-ELISA binding
11	Dfarinae09015	Heat shock protein P90	718	Asp f 12 (AAB51544) 65.34% <i>Aspergillus fumigatus</i>	None tested
12	Dfarinae08265	Aldolase A	361	Pan h 3 (XP_026771637.1) 65.00% <i>Pangasianodon hypophthalmus</i>	None tested
13	Dfarinae07232	Pyruvate kinase	533	Sal s 9 (ACH70965.1) 63.28% <i>Salmo salar</i>	None tested
14	Dfarinae12793	Aldehyde dehydrogenase	451	Tyr p 35 (AOD75396.1) 62.94% <i>Tyrophagus putrescentiae</i>	None tested
15	Dfarinae11953	Peptidyl-prolyl cis-trans isomerase protein	227	Asp f 27 (CAI78448) 59.15% <i>Aspergillus fumigatus</i>	No IgE-ELISA binding
16	Dfarinae03857	Troponin I	211	Pon l 7 (P05547) 57.31% <i>Pontastacus leptodactylus</i>	None tested
17	Dfarinae09176	Pyruvate kinase PKM-like	528	Pan h 9 (XP_026775867.1) 57.17% <i>Pangasianodon hypophthalmus</i>	None tested
18	Dfarinae10217	Eukaryotic translation initiation factor 3 subunit	328	For t 2 (ACD65081) 56.13% <i>Forcipomyia taiwana</i>	None tested
19	Dfarinae04475	Transaldolase	327	Pen ch 35 (ADK27483) 56.04% <i>Penicillium chrysogenum</i>	None tested
20	Dfarinae04273	Peptidyl-prolyl isomerase	209	Asp f 11	None tested

				(CAB44442) 55.84% <i>Aspergillus fumigatus</i>	
21	Dfarinae05803	manganese superoxide dismutase	231	Pis v 4 (ABR29644) 55.45% <i>Pistacia vera</i>	None tested
22	Dfarinae09701	Mitochondrial malate dehydrogenase	343	Mala f 4 (AAD25927) 54.68% <i>Malassezia furfur</i>	None tested
23	Dfarinae05230	Myosin, light chain	202	Bla g 8 (ABD47458) 54.6% <i>Blattella germanica</i>	None tested
24	Dfarinae11869	Lysosomal aspartic protease protein	401	Aed a 11 (XP_001657556.1) 54.36% <i>Aedes aegypti</i>	No IgE-ELISA binding
25	Dfarinae03060	Elongation factor 1	73	Tri a 45 (AKJ77985.1) 52.17% <i>Triticum aestivum</i>	None tested
26	Dfarinae02193	Thioredoxin	106	Plo i 2 (CBW45298) 51.96% <i>Plodia interpunctella</i>	None tested
27	Dfarinae01264	ML-domain protein	143	Blo t 2 (AAQ73483) 50.69% <i>Blomia tropicalis</i>	None tested
28	Dfarinae12442	Aldehyde dehydrogenase protein	490	Cla h 10 (CAA55072) 50.63% <i>Cladosporium herbarum</i>	No IgE-ELISA binding
29	Dfarinae04109	Acid ribosomal protein P1	119	Cla h 12 (CAA59463) 50% <i>Cladosporium herbarum</i>	None tested

TABLE S13.

Location	Allergen	Locus tag*	Biochemical function	Deduced no. amino acids	WHO/IUIS data [†] (% similarity)
Contig 1	Der f 1	Dfarinae09850	Cysteine protease	321	BAC53948.1 (100.00%)
	Der f 8	Dfarinae08204	Glutathione S-transferase	217	AGC56215.1 (100.00%)
	Der f 18	Dfarinae08412	Chitin-binding protein	462	AAM19082.1 (100.00%)
Contig 2	Der f 3	Dfarinae05328	Trypsin	258	AAP35076.1 (99.22%)
	Der f 5	Dfarinae04701	Low molecular weight IgE binding protein	132	ABO84970.1 (100.00%)
	Der f 21	Dfarinae04702	SH3 domain-binding glutamic acid-rich protein	136	ABO84964.1 (99.26%)
	Der f 27	Dfarinae03993	Serpin	427	AIO08851.1 (99.77%)
	Der f 33	Dfarinae04694	Alpha-tubulin	461	AIO08861.1 (99.78%)
Contig 3	Der f 2	Dfarinae00310	NPC2 family	146	ABG76194.1 (97.95%)
	Der f 13	Dfarinae00170	Fatty acid binding protein	131	AAP35078.1 (100.00%)
	Der f 14	Dfarinae0323	Apolipoprotein	1631	AAM21322.1 (86.34%)
	Der f 25	Dfarinae01490	Triosephosphate isomerase	279	AIO08860.1 (100.00%)
	Der f 28	Dfarinae01557	Heat Shock Protein	654	AIO08848.1 (100.00%)
	Der f 30	Dfarinae00321	Ferritin	166	AGC56219.1 (96.79%)
	Der f 31	Dfarinae01441	Cofilin	148	AIO08870.1 (100.00%)
	Der f 35	Dfarinae01264	Uncharacterized protein	143	BAX34757.1 (100.00%)
Contig 4	Der f 6	Dfarinae12003	Chymotrypsin	279	ABG23667.1 (99.64%)
	Der f 7	Dfarinae11783	Bactericidal permeability-increasing like protein	213	ACK76298.1 (99.53%)
	Der f 10	Dfarinae11789	Tropomyosin	284	ABU97468.2 (100.00%)
	Der f 22	Dfarinae10875	lipid binding protein	155	ABG35122.1 (100.00%)
Contig 5	Der f 4	Dfarinae06737	Alpha-amylase	525	AJF93907.1 (99.81%)
	Der f 37	Dfarinae06505	Chitin binding protein	250	QBF67839.1 (100.00%)
Contig 6	Der f 9	Dfarinae02538	Collagenolytic serine protease	276	AIO08869.1 (98.55%)
	Der f 15	Dfarinae02404	Chitinase	573	AAD52672.1 (96.68%)
	Der f 36	Dfarinae02711	Uncharacterized protein	229	ATI08931.1 (100.00%)
Contig 7	Der f 16	Dfarinae05870	Gelsolin/villin	480	AAM64112.1 (100.00%)
	Der f 23	DfarinaeADD06015	Peritrophin-like protein	174	ALU66112.1 (98.28%)
	Der f 29	Dfarinae06374	Peptidyl-prolyl cis-trans isomerase (cyclophilin)	306	AAP35065.1 (100.00%)
	Der f 32	Dfarinae05847	Secreted inorganic pyrophosphatase	296	AIO08849.1 (100.00%)
	Der f 34	Dfarinae06301	Enamine/imine deaminase	164	A0A1J1DL12.1 (99.22%)
Contig 8	Der f 39	Dfarinae12240	Troponin C	153	QBF67841.1 (100.00%)

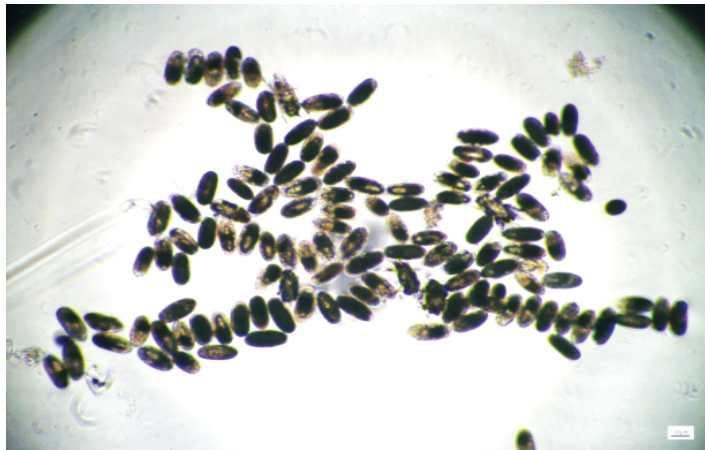
Contig 9	Der f 11	Dfarinae07639	Paramyosin	875	AIO08864.1 (99.89%)
	Der f 20	Dfarinae07890	Arginine kinase	311	AIO08850.1 (99.68%)
	Der f 24	Dfarinae07745	Ubiquinol-cytochrome c reductase binding protein	118	AGI78542.1 (100.00%)
	Der f 26	Dfarinae07770	Myosin alkali light chain	160	ATI08939.1 (93.12%)

*Locus tags and contig number are included in our assembled *D. farinae* genome.

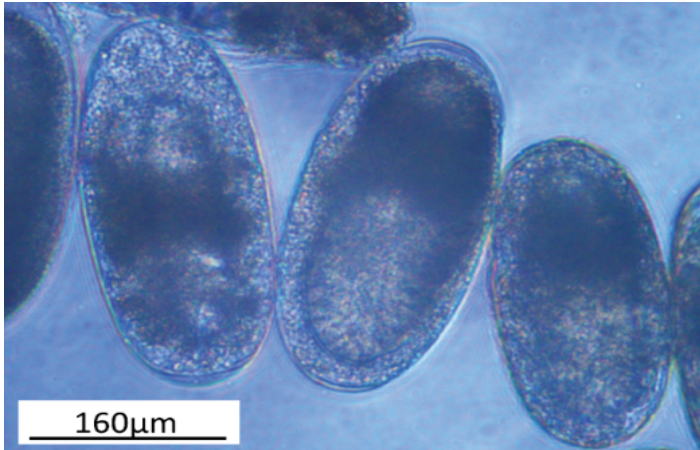
†WHO/IUIS Der f 17 listing has no sequence information. There are no Group 12 or 19 allergens in Der f. The bacterial lytic enzyme like protein (GenBank accession no. MT360919) in Der f (Der p 38 homolog) had no IgE-binding activity and was not confirmed as an allergen in this study. Analyzed relative to the NCBI database by BLAST algorithm; GenBank accession numbers are listed. Der f 9 (AIO08869.1) protein sequence is not listed in WHO/IUIS website. Der f 14 (GenBank accession no. BAA04558) has partial sequence.

Figure S1

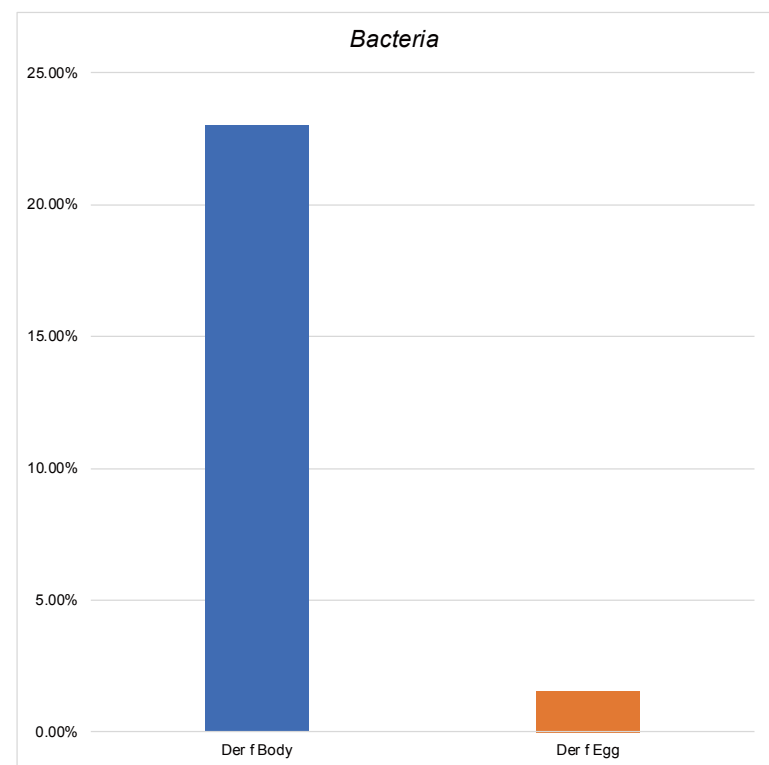
A



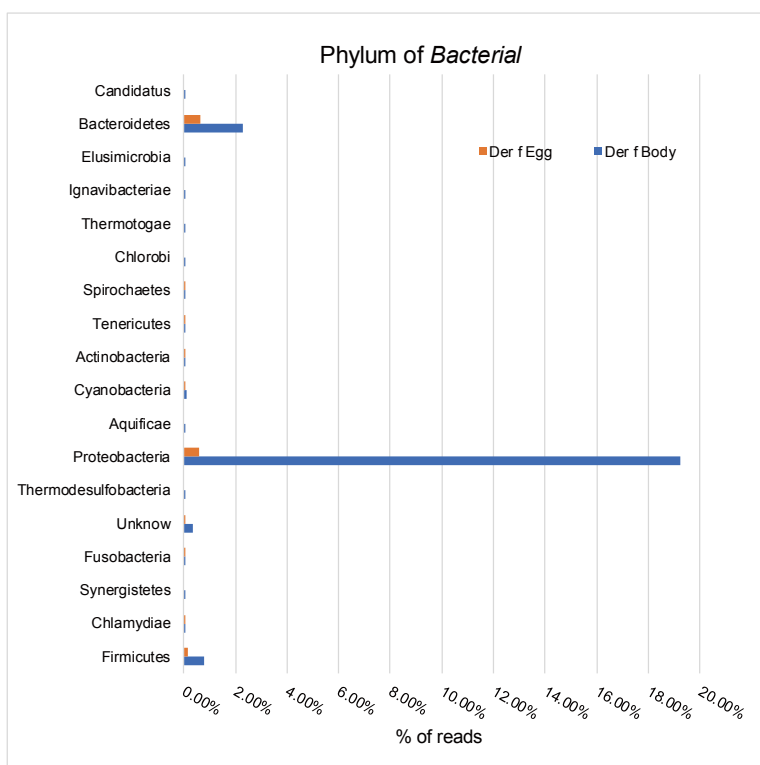
B



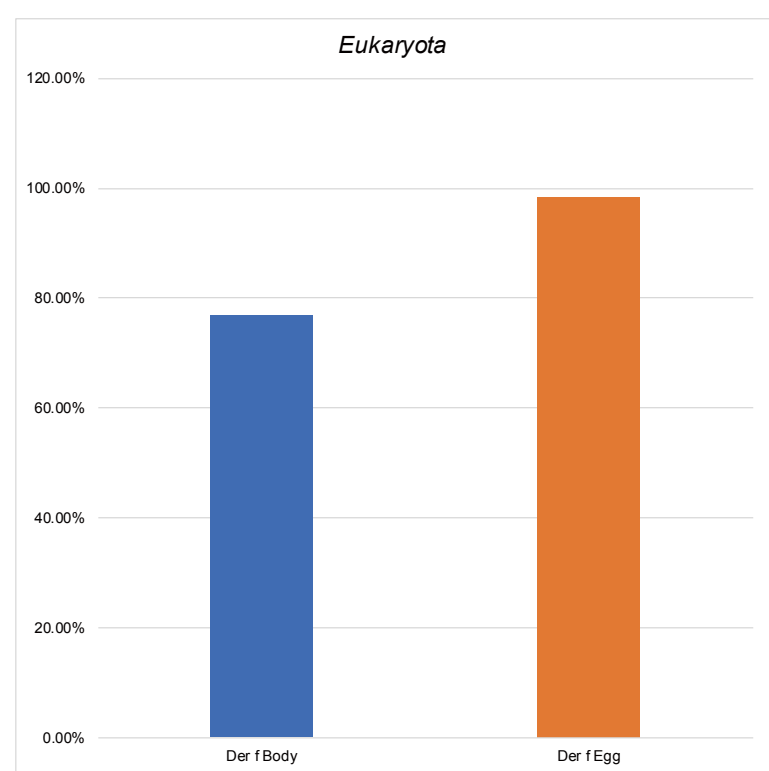
C



D



E



F

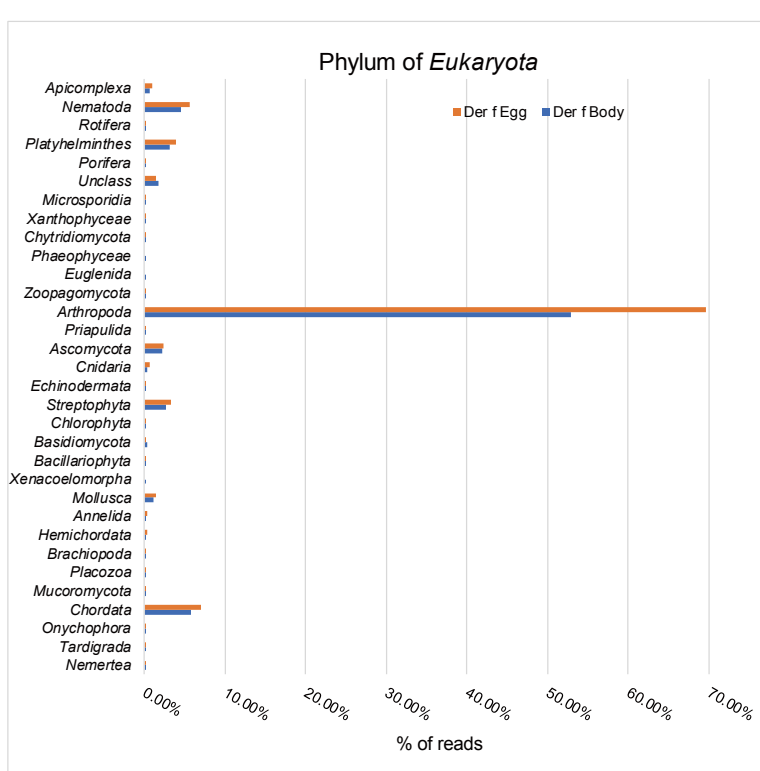


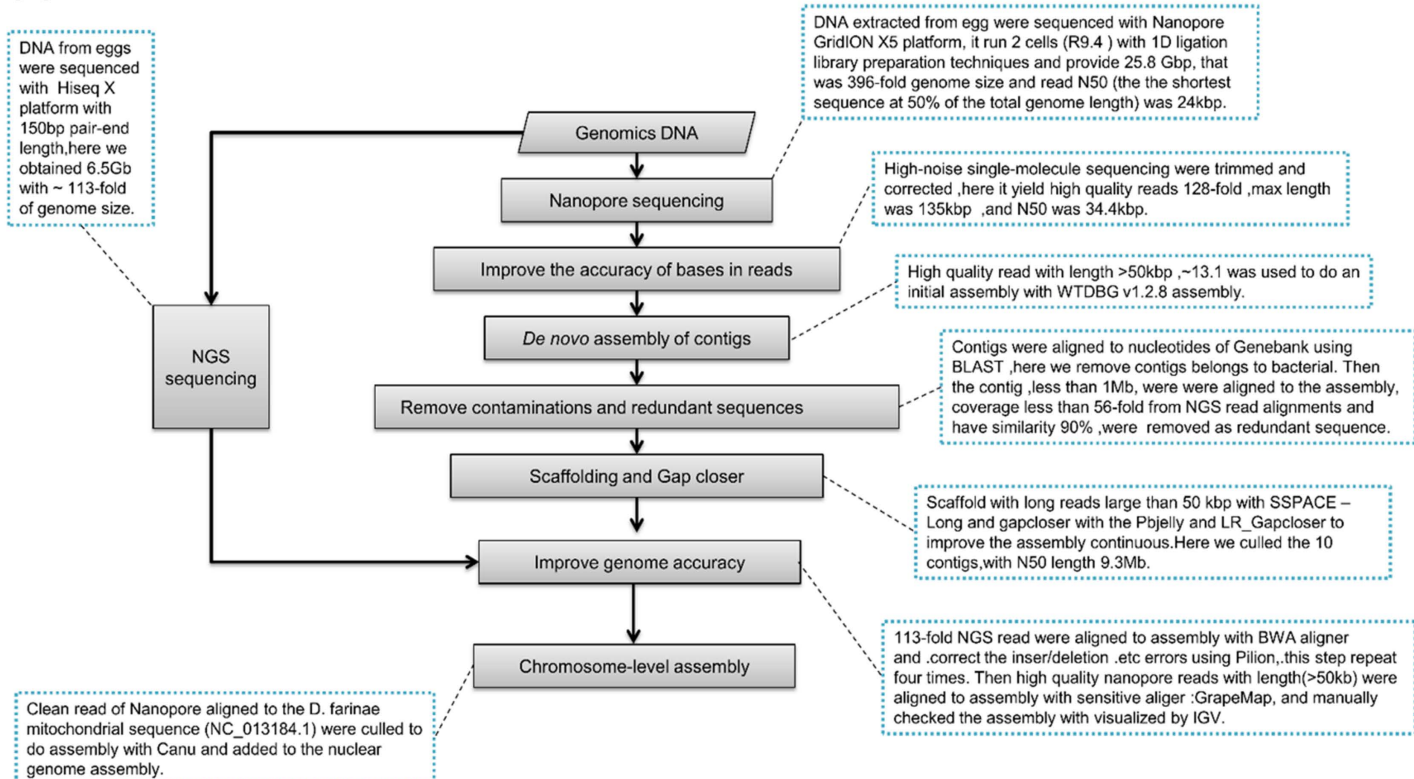
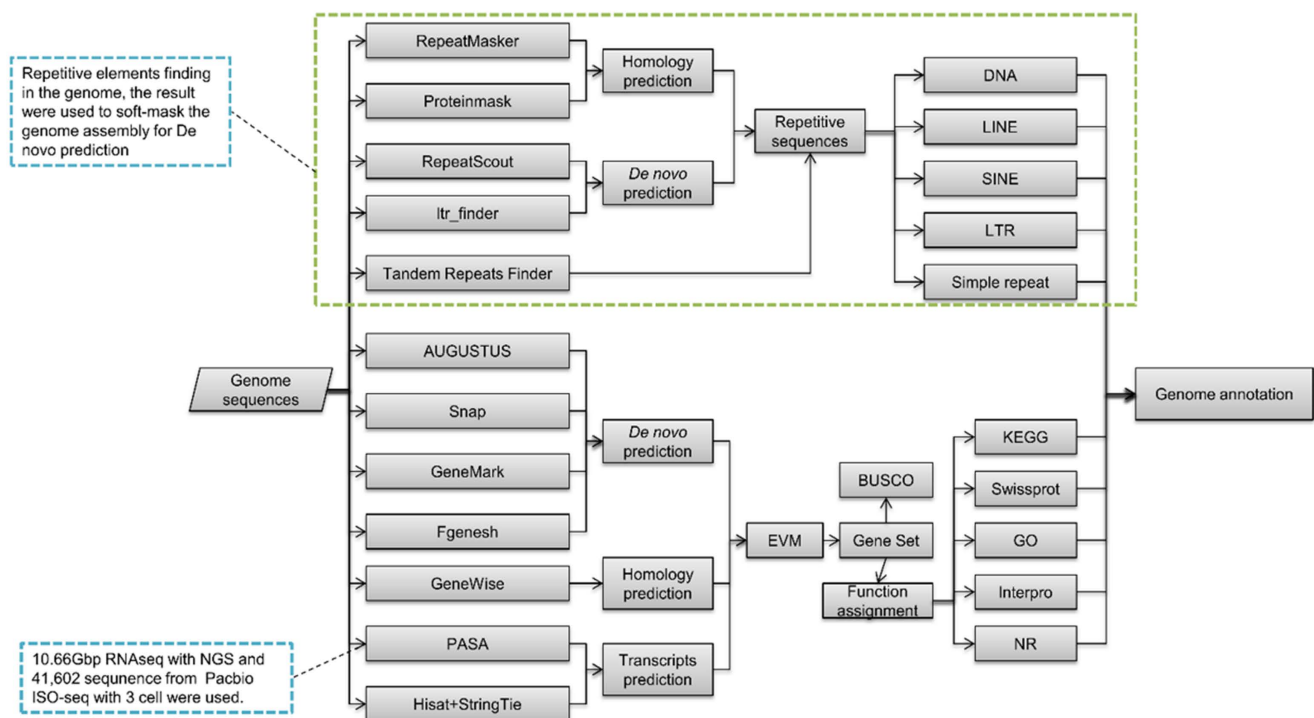
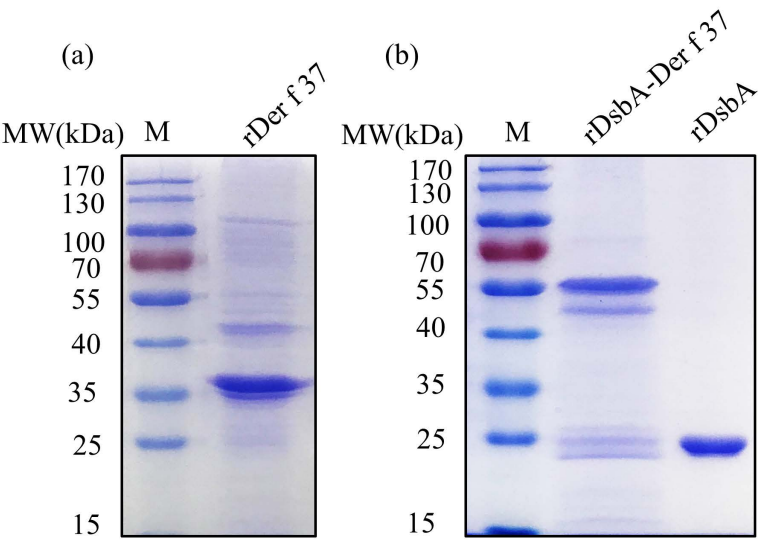
Figure S2**A****B**

Figure S3

A

001 MIKFYFLFCSFYIVSVWMLPVSELAGMHQLDEGMINVDDNKPTNLTICVD
051 KKDGTLLGDEHDRTLIFYECSNGLAFPFHCPSNLIFDETRGVCFDPNSPT
101 KRPPNHGNSTYNTCTGKEDGFYPDENDATKFHECVNGYPYDFKCPPNTIYD
151 VKRKVCAYKSIERPKEENEQPQPQPQQQQHRAVRDVIESTTAEIKEELPT
201 EQSVRRHRRNDEQQQQQHHHHQLEEFakeKNIDLHIIETTTQENFDVWID

B



C

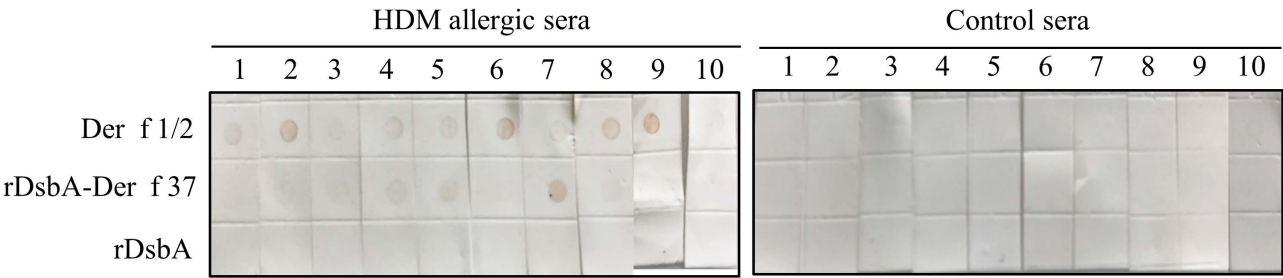
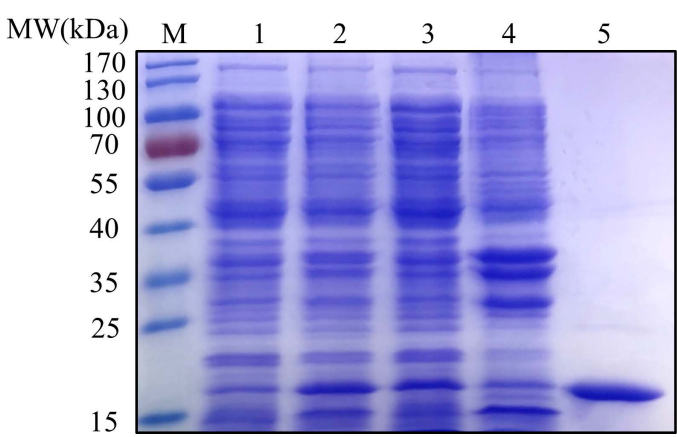


Figure S4

A

001 MSVEELTKEQVQMLRKAFDMFDREKKGHIHTNMVSTILRTLGGQTFEEKDL
051 QQLITEIDADGSGELEFDEFLLTLARFLVEEDAEAMQEELREAFRMYDKE
101 GNGYIPTSALREILRALDDKLTNDELDEMIAEIDTDGSGTVDFDEFMEMM
151 TGD

B



C

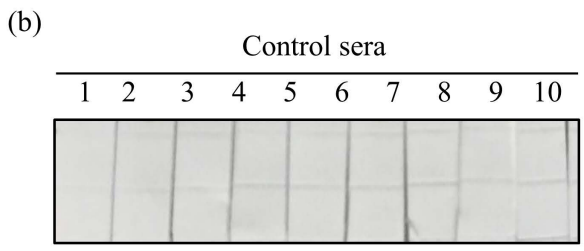
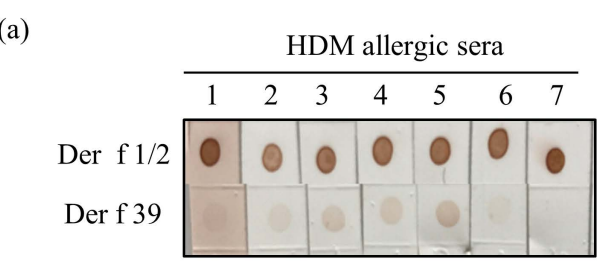


Figure S5

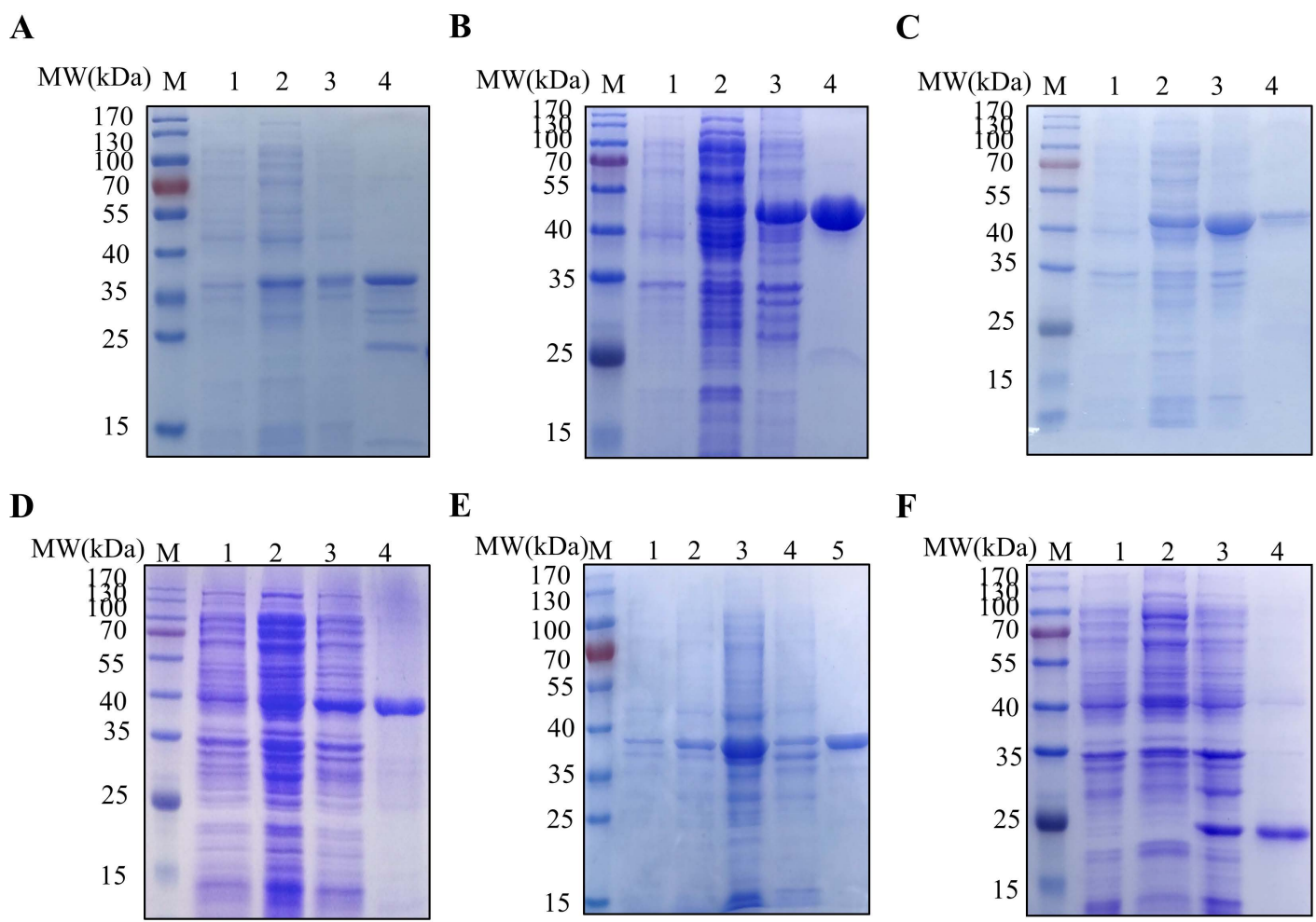


Figure S6

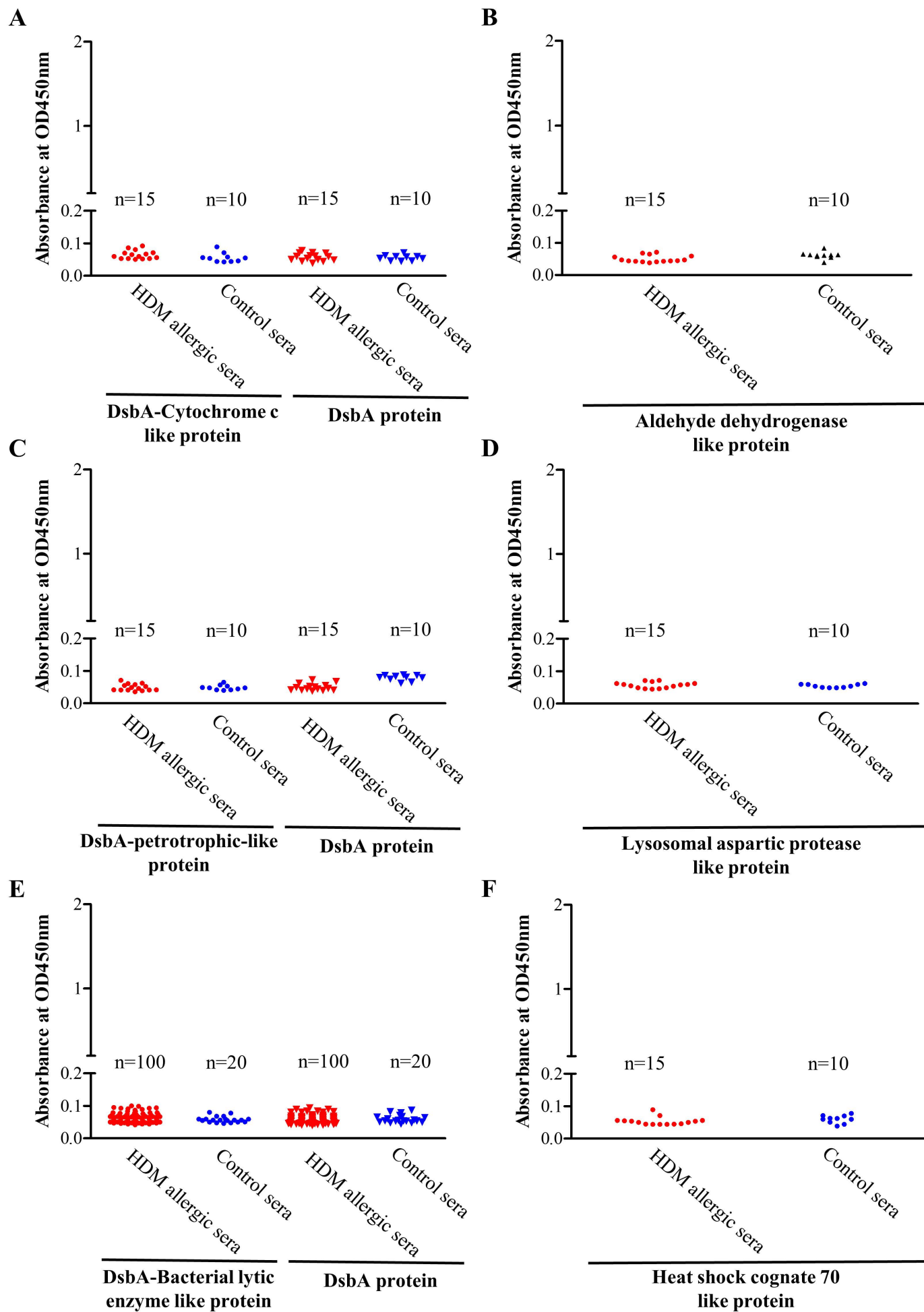


Figure S7

MT360919	MKFFLTLTLFCALAINHVGDGSHIVKAARSQIGVPYSWGGGGIHGKSKGIGEGANIVGF	60
QH72282.1	MKLFLTLTLFCALAIHVYGDGSHIVKAARSQIGVPYSWGGGGIHGKSKGIGEGANIVGF	60
	** : *****. *****	
MT360919	DCSGLAQYSIYQGTHKTIARTAAQYNDNHCHHVAYGSHQPGDLVFFGNPIYHVGIVSAH	120
QH72282.1	DCSGLAQYSIYQGTHKTIARTAAQYNDNHCHHVAYGSHQPGDLVFFGNPIYHVGIVSAH	120

MT360919	GRMVNAPKPGTKVREENIWSYHISHVARCW	150
QH72282.1	GRMVNAPKPGTKVREENIWSYHISHVARCW	150
