



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

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

## 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.<sup>1-4</sup> In-depth analyses of the full spectrum of HDM allergens are needed to elucidate HDM allergy mechanisms and guide diagnostic and immunotherapeutic development.<sup>5</sup>

Publication of a Der f draft genome produced with next generation sequencing (NGS) facilitated the identification of numerous allergens,<sup>6-11</sup>

<sup>a</sup>Department 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](mailto:sunjinlv@pumch.cn) or [Jiang.Lee2018@gmail.com](mailto:Jiang.Lee2018@gmail.com) or [jkm@szu.edu.cn](mailto: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).<sup>12</sup> However, the Der f draft genome has shortcomings due to technical limitations.<sup>5</sup> For example, the Der f 23 cDNA sequence differs from its draft genome corollary.<sup>13</sup> Because microbiota sequences were removed manually, the draft may contain microbiome sequences.<sup>5</sup> Additionally, limitations of short-read sequencing were likely to produce much scaffold gaps.<sup>5</sup>

During genome assembly, it is important to minimize cross-species DNA contamination.<sup>14</sup> 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.

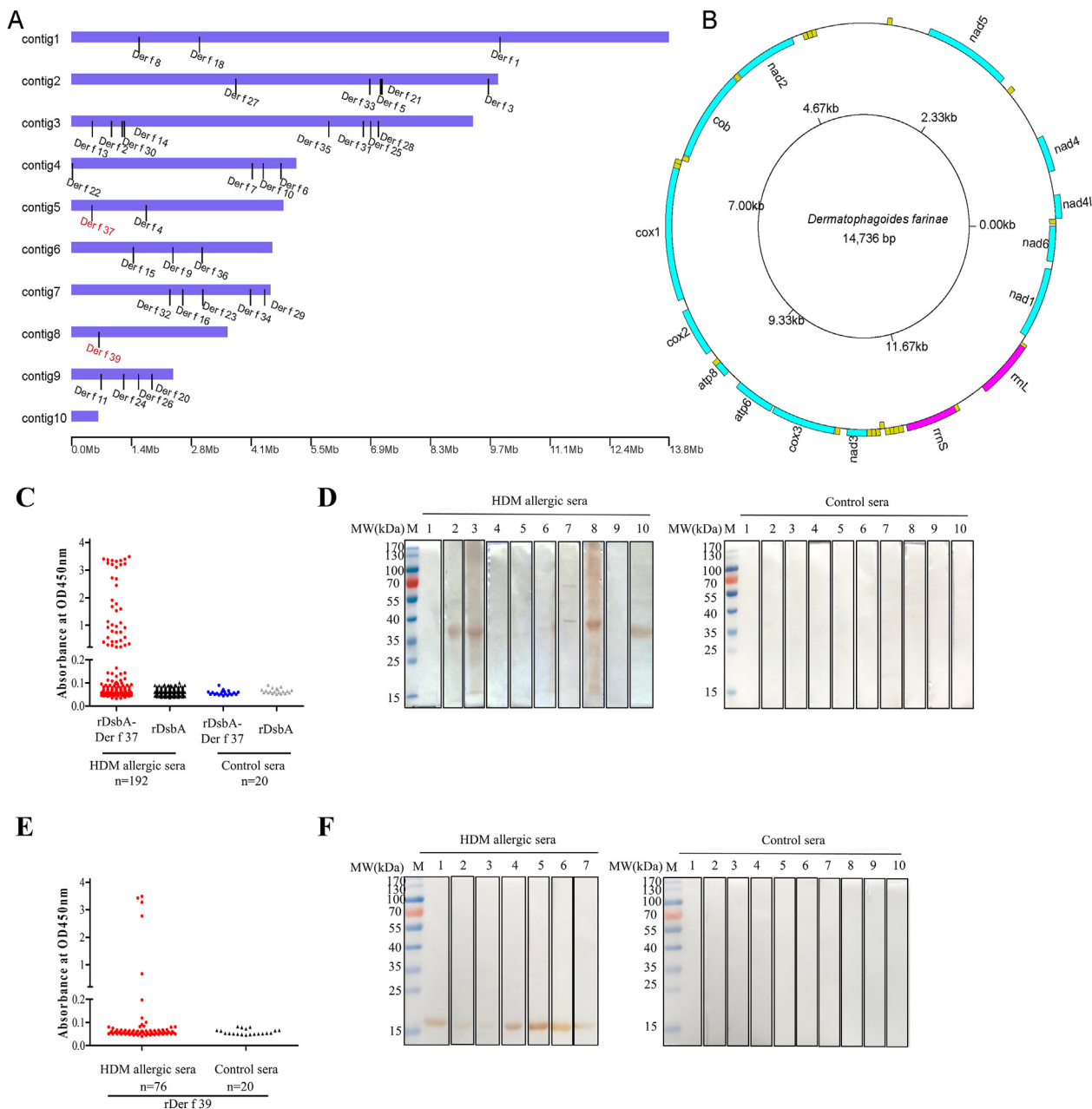
Due to the symbiotic relationship in the digestive tract, it was not possible to obtain pure mite bodies without any microorganisms through aseptic culture methods.<sup>5</sup> Our attempts to isolate Der f cells aseptically were not successful (data not shown). To minimize microbial DNA contamination, 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<sub>A</sub>/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.<sup>13</sup> 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).<sup>15</sup>

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.<sup>16</sup> 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 ( $N = 15$ ;  $>100 \text{ kU}_A/\text{L}$ ), 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 <sup>b</sup>
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 <sup>a</sup> 100% (7/7); IgE-dot blot <sup>a</sup> 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. <sup>a</sup>HDM allergic sera from positive samples in IgE-ELISA assay. <sup>b</sup>WHO/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

<sup>a</sup>Department of Biochemistry and Molecular Biology, Laboratory Department of Pinghu Hospital, Health Science Center, Shenzhen University, Shenzhen, 518060, China  
<sup>b</sup>EasyATCG L.L.C, Shenzhen, 518000, China <sup>c</sup>Shenzhen University General Hospital, Shenzhen, 518060, China <sup>d</sup>Allergy 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.