The draft genome, transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens

Ting-Fung Chan, PhD,[a]* Kun-Mei Ji, PhD,[b]* Aldrin Kay-Yuen Yim, BS,[a,c]* Xiao-Yu Liu, MSc,[d]* Jun-Wei Zhou, PhD,[e]*‡ Rui-Qi Li, BSM,[d]* Kevin Yi Yang, PhD,[c,e]* Jing Li, MS,[f]* Meng Li, MSc,[d] Patrick Tik-Wan Law, PhD,[a] Yu-Lan Wu, BS,[d] Ze-Lang Cai, MSc,[d] Hao Qin, PhD,[a,c] Ying Bao, MSc,[d] Ross Ka-Kit Leung, PhD,[c,e] Patrick Kwok-Shing Ng, PhD,[e] Ju Zou, MSc,[d] Xiao-Jun Zhong, MSc,[d] Pi-Xin Ran, MD,[f] Nan-Shan Zhong, MSc,[f] Zhi-Gang Liu, MD,[b,d] and Stephen Kwok-Wing Tsui, PhD[c,e] *Hong Kong, Shenzhen, and Guangzhou, China*

Background: A sequenced house dust mite (HDM) genome would advance our understanding of HDM allergens, a common cause of human allergies.

Objective: We sought to produce an annotated *Dermatophagoides farinae* draft genome and develop a combined genomic-transcriptomic-proteomic approach for elucidation of HDM allergens.

Methods: A *D farinae* draft genome and transcriptome were assembled with high-throughput sequencing, accommodating microbiome sequences. The allergen gene structures were validated by means of Sanger sequencing. The mite's microbiome composition was determined, and the predominant genus was validated immunohistochemically. The allergenicity of a ubiquinol-cytochrome c reductase binding protein homologue was evaluated with immunoblotting, immunosorbent assays, and skin prick tests.

Results: The full gene structures of 20 canonical allergens and 7 noncanonical allergen homologues were produced. A novel major allergen, ubiquinol-cytochrome c reductase binding protein–like protein, was found and designated Der f 24. All 40 sera samples from patients with mite allergy had IgE antibodies against rDer f 24. Of 10 patients tested, 5 had positive skin reactions. The predominant bacterial genus among 100 identified species was *Enterobacter* (63.4%). An intron was found in the 13.8-kDa *D farinae* bacteriolytic enzyme gene, indicating that it is of HDM origin. The Kyoto Encyclopedia of Genes and Genomes pathway analysis revealed a phototransduction pathway in *D farinae*, as well as thiamine and amino acid synthesis pathways, which is suggestive of an endosymbiotic relationship between *D farinae* and its microbiome.

Conclusion: An HDM genome draft produced from genomic, transcriptomic, and proteomic experiments revealed allergen genes and a diverse endosymbiotic microbiome, providing a tool for further identification and characterization of HDM allergens and development of diagnostics and immunotherapeutic vaccines. (J Allergy Clin Immunol 2015;135:539-48.)

***Key words:** House dust mite, allergen, genome, microbiome, transcriptome, proteome, ubiquinol-cytochrome c reductase binding protein, Der f 24,* Enterobacter *species*

Allergic diseases, which affect 30% to 40% of the world's population and are increasing in prevalence internationally, particularly among young people, have negative effects on patients' work and social lives and have become a costly global health problem.[1,2] House dust mites (HDMs) are predominant sources of inhalant allergens, with more than 50% of allergic disease cases being attributed to them.[3-5] Decades of research have revealed 23 HDM allergen groups, with the canonical group 1 and 2 allergens being the most clinically important because they possess IgE-binding activity in most sera of patients with mite allergy.[5-7] Group 1 and 2 allergens induce T_H2 immune responses by encoding cysteine proteases and by facilitating Toll-like receptor 4 signaling, respectively.[8,9]

It remains a perplexing question why HDMs are seemingly teeming with allergenic components. The identities of the full spectrum of HDM allergenic components are not yet known. Allergen-specific immunotherapy represents the only currently available therapy that has long-lasting effects on allergic diseases.[10] HDM allergen vaccines are generally made from extracts of purified mite bodies, which include components of microbes that inhabit mites.[11,12] It is difficult to ensure the lot-lot consistency of the vaccine because of its complex components. Distinguishing the effective components of vaccines

From [a]the School of Life Sciences, [c]the Hong Kong Bioinformatics Centre, and [e]the School of Biomedical Sciences, Chinese University of Hong Kong; [b]the State Key Laboratory of Respiratory Disease for Allergy at Shenzhen University and [d]the Shenzhen Key Laboratory of Allergy and Immunology, School of Medicine, Shenzhen University; and [f]the State Key Laboratory of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical University.

*These authors contributed equally to this work.

‡Jun-Wei Zhou, PhD, is currently affiliated with the Basic Medical College of Zhengzhou University, Zhengzhou, China.

Supported by the NSFC (81071388; 31328014; 3027122), Guangdong-Hong Kong Technology Cooperation Foundation (2005A10905007), the China 863 (2006AA02A231) and 973 (2009CB522100) Project Foundations, the Chinese University of Hong Kong Focused Investment Funding, the Hong Kong Research Grants Council (475113), the Shenzhen Science Foundation (JCYJ20120613113021045; JCYJ20130326112225593) and the Guangdong Province NSF (no. 2014566).

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

Received for publication June 29, 2013; revised September 3, 2014; accepted for publication September 25, 2014.

Available online October 31, 2014.

Corresponding author: Nan-Shan Zhong, MSc, State Key Laboratory of Respiratory Disease, the First Affiliated Hospital of Guangzhou Medical College No. 151, Yanjiang Rd, Guangzhou 510120, China. E-mail: nanshan@vip.163.com. Or: Zhi-Gang Liu, MD, State Key Laboratory of Respiratory Disease for Allergy at Shenzhen University and Shenzhen Key Laboratory of Allergy and Immunology, School of Medicine, Shenzhen University, Nanhai Ave 3688, 518060 Shenzhen, Guangdong, People's Republic of China. E-mail: lzg@szu.edu.cn. Or: Stephen Kwok-Wing Tsui, PhD, School of Biomedical Sciences, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: kwtsui@cuhk.edu.hk.

0091-6749

http://dx.doi.org/10.1016/j.jaci.2014.09.031

Abbreviations used

CEG:	Core eukaryotic gene
CEGMA:	Core Eukaryotic Genes Mapping Approach
Gb:	Gigabase
GO:	Gene ontology
HDM:	House dust mite
KEGG:	Kyoto Encyclopedia of Genes and Genomes
NCBI:	National Center for Biotechnology Information
UQCRB:	Ubiquinol-cytochrome c reductase binding protein

from those that produce side effects would enable more potent and safe vaccines to be developed.

Having knowledge of the HDM genome and its endosymbiotic microbiome will be pivotal to resolving the aforementioned core scientific and clinical issues in the field of allergy. The closest species to the HDM for which a genome draft has been produced is the spider mite *Tetranychus urticae*,[13] which is a cause of occupational allergic disease in agricultural workers.[14] However, despite their prominent role as allergen sources, the genomes of the HDMs *Dermatophagoides pteronyssinus* and *Dermatophagoides farinae* have yet to be resolved, restricting more in-depth research on HDM allergens and the mechanisms underlying their allergenicity. Here, we combined genomic and transcriptomic approaches to produce a *D farinae* draft genome that can provide insights into the identities of the full array of *D farinae* allergens and the mechanisms mediating their allergenicity, including the potential role of the microbiome. We applied our draft genome in combination with proteomic and comparative analyses to uncover a novel major allergen and examine the genes underlying physiologic and metabolic processes.

METHODS

Mite culture and purity check

D farinae mites were isolated from indoor dust samples from Shenzhen City in southern China.[2] The mite culture and purity check methods are described in the Methods section in this article's Online Repository at www.jacionline.org.

Genome and transcriptome sequencing

D farinae genomic DNA and RNA samples were prepared as described in the Methods section in this article's Online Repository. Four paired-end sequencing libraries with insert sizes of 200, 500, 2000, and 5000 bp, respectively, were constructed by using *D farinae* whole DNA and then sequenced with an Illumina HiSeq 2000 Sequencer. A total of 24 gigabase (Gb) pairs of sequencing data were generated. *D farinae* cDNAs were sequenced with the Illumina HiSeq 2000 Sequencer; 5.8 Gb of paired-end sequencing data (insert size, approximately 200 bp) was generated for transcriptome analysis.

Genome assembly and annotation using transcriptome data

Genome assembly began with reconstruction by using SOAPdenovo,[15] ALLPATHS-LG,[16] and Velvet[17] (see the Methods section in this article's Online Repository). Protein-coding genes were predicted with the use of 2 *ab initio* gene prediction tools: GeneMark-ES[18] and GimmerHMM.[19] Annotation of noncoding RNA genes was done with tRNAscan-SE[20] and RNAmmer.[21] Transcriptome sequencing data were assembled by using Trinity,[22] and the assembled transcripts were used to refine the annotations by using GeneMark-ES and GlimmerHMM. Splice junctions and relative abundance of RNA sequencing reads were determined with TopHat,[23] SpliceMap,[24] and Cufflink.[25] Finally, we evaluated the completeness of our draft genome relative to the Core Eukaryotic Genes Mapping Approach (CEGMA) set of 248 core eukaryotic genes (CEGs) with the CEGMA pipeline.[26]

Microbiome analysis

Because of the possibility of symbiotic relationships between mites and microorganisms that might preclude entirely sterile culture conditions, mite sequencing data were separated from microbiota sequences by means of manual curation based on BLAST searches in the microbial database. The assembled draft genome was compared with microbial databases, as described in detail in the Methods section in this article's Online Repository, to distinguish between *D farinae* and microbial genomes. Briefly, we searched the microbial RefSeq database using genomic sequencing reads with a high-stringency cutoff (E-value $\leq 1e^{-50}$); matches for each read must map to the same genus.

The following aspects are described in the Methods section in this article's Online Repository: mite culture and purity check; Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and phylogenetic analyses and metabolic comparison to *Tetranychus urticae*; allergen gene cloning and proteomic identification; IgE-binding assay and skin prick tests; and *Enterobacter cloacae* immunohistochemistry and bacteriolytic enzyme gene cloning.

RESULTS

Mite culture purity

Morphologic inspections and PCR experiments confirmed the purity of the *D farinae* species in our cultures. There was no contamination from *D pteronyssinus*. After the culture medium had been digested with nuclease, the genomic DNA sample was confirmed to be contamination free (Fig 1, *A* and *B*, and see Figs E1-E3 and Tables E1 and E2 in this article's Online Repository at www.jacionline.org).

Mite genome draft

High-throughput sequencing (see Table E3 in this article's Online Repository at www.jacionline.org) yielded 24 Gb of genome sequences or roughly 380-fold coverage of the estimated genome size. After building a *de novo* draft assembly and applying gap filling, 4 sequencing libraries were assembled into 554 scaffolds (total length, 61 Mb; N50 length, 197 kb). Because the sample included nucleic acids attributable to the mite's microbiome, we examined its microbial composition (see Table E4 in this article's Online Repository at www.jacionline.org). Separation of microbial DNA resulted in a 53.5-Mb *D farinae* draft genome with 516 nuclear genome scaffolds (N50 = 187 kb) and a 14.3-kb mitochondrial genome (see Table E5 in this article's Online Repository at www.jacionline.org). The draft genome was submitted to the National Center for Biotechnology Information (NCBI) BioProject (ID: PRJNA17406, accession no.: ASGP00000000).

Our draft genome included 242 (97.58%) of 248 CEGs, with 239 (96.58%) of 248 complete CEGs (see Table E6 in this article's Online Repository at www.jacionline.org), indicating good completeness.[22] We retrieved 264 nucleotide sequences and 189 amino acid sequences of *D farinae* from the NCBI (April 2012) and confirmed that 261 (98.8%) of the nucleotide sequences and 182 (96.3%) of the amino acid sequences were present in the draft genome (E-value cutoff: $1e^{-6}$).

The guanine-cytosine contents of the coding DNA sequences and whole genome were 34.4% and 29.5%, respectively.

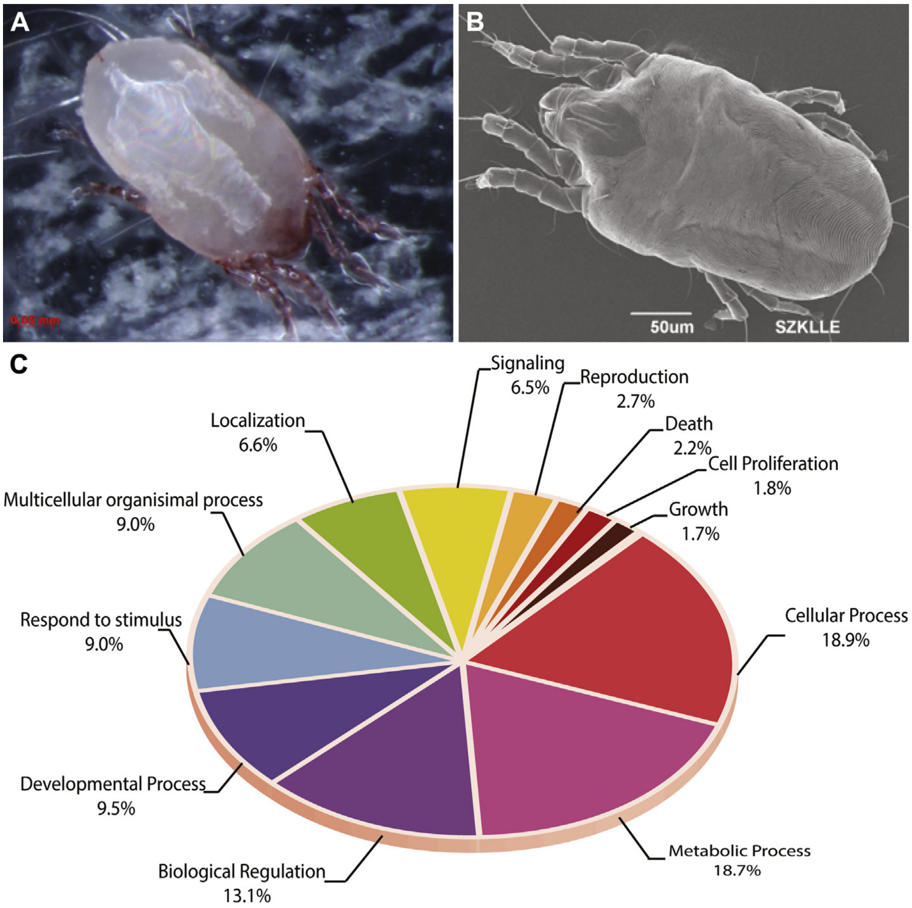

FIG 1. HDM morphology and phylogeny. **A,** Photomicrograph of a live adult female *D farinae* mite. **B,** Scanning electron microscopic image of an adult female *D farinae* mite. **C,** GO distribution of *D farinae*. **D**, Phylogenetic tree of *D farinae* with 25 other arthropod species.

GeneMark-ES, GimmerHMM run 1 (with *Caenorhabditis elegans*), and GimmerHMM run 2 (with *T urticae*) predicted 13,475, 20,165, and 14,156 gene models, respectively. Another 3,265 gene models were inferred with TopHat and Cufflinks transcriptome sequence analysis. Altogether, 16,376 gene models were obtained, including 9,142 that were supported by RNA sequencing results. Gene annotation with Blast2GO yielded 8,201 genes with significant hits (E-value cutoff: $1e^{-10}$) in the NCBI nonredundant database. Among them, 7,348 were assigned at least 1 gene ontology (GO) term (Fig 1, *C*).

Using tRNAscan-SE, we identified 65 annotated transfer RNA genes within the *D farinae* nuclear genome. The rRNA genes (18S, 5.8S, and 28S) that were experimentally identified in *D farinae* previously were found within our *D farinae* genome draft (see Table E5). Both 18S and 28S were predicted in their entirety by using RNAmmer.

Phylogenetic analysis

A phylogenetic tree was constructed from the protein sequences of 101 CEGs that are shared among 25 Arachnida, Branchiopoda, and Insecta species (see Table E7 in this article's Online Repository at www.jacionline.org) to reveal *D farinae*'s phylogenetic relationships with other arthropods. A comparison of *D farinae*'s genome annotations with those of *T urticae*, to which our maximum likelihood tree showed *D farinae* to be phylogenetically close (Fig 1, *D*), revealed similar GO distributions (see Fig E4 in this article's Online Repository at www.jacionline.org).[23] Regarding genes associated with metabolic processes, 662 of 3,029 genes in *D farinae* and 476 of 2,492 genes in *T urticae* did not correlate with each other.

Transcriptome and KEGG pathway analysis

We assembled 5.8 Gb of paired-end sequencing data, annotating a total of 16,376 genes, with only 9,142 (55.8%) being supported by RNA sequencing data, probably because of developmental variation of gene expression profiles. Functional annotation for GO and the KEGG[27] pathway database using the BLAST2GO[28] program mapped 7,348 protein-coding genes to GO project categories (Fig 1, *C*). We identified most constituents of the KEGG phototransduction pathway in the *D farinae* annotated genes (see Fig E5, *A*, in this article's Online Repository at www.jacionline.org) with high homology, with the exception of rhodopsin. One candidate gene (DEFA_098690) encoding a 7-transmembrane domain protein with 83.7% similarity to the rhodopsin family transmembrane receptor domain (aa 84–206; XP_002430048 in GenBank) of *Pediculus humanus corporis* was annotated as a class

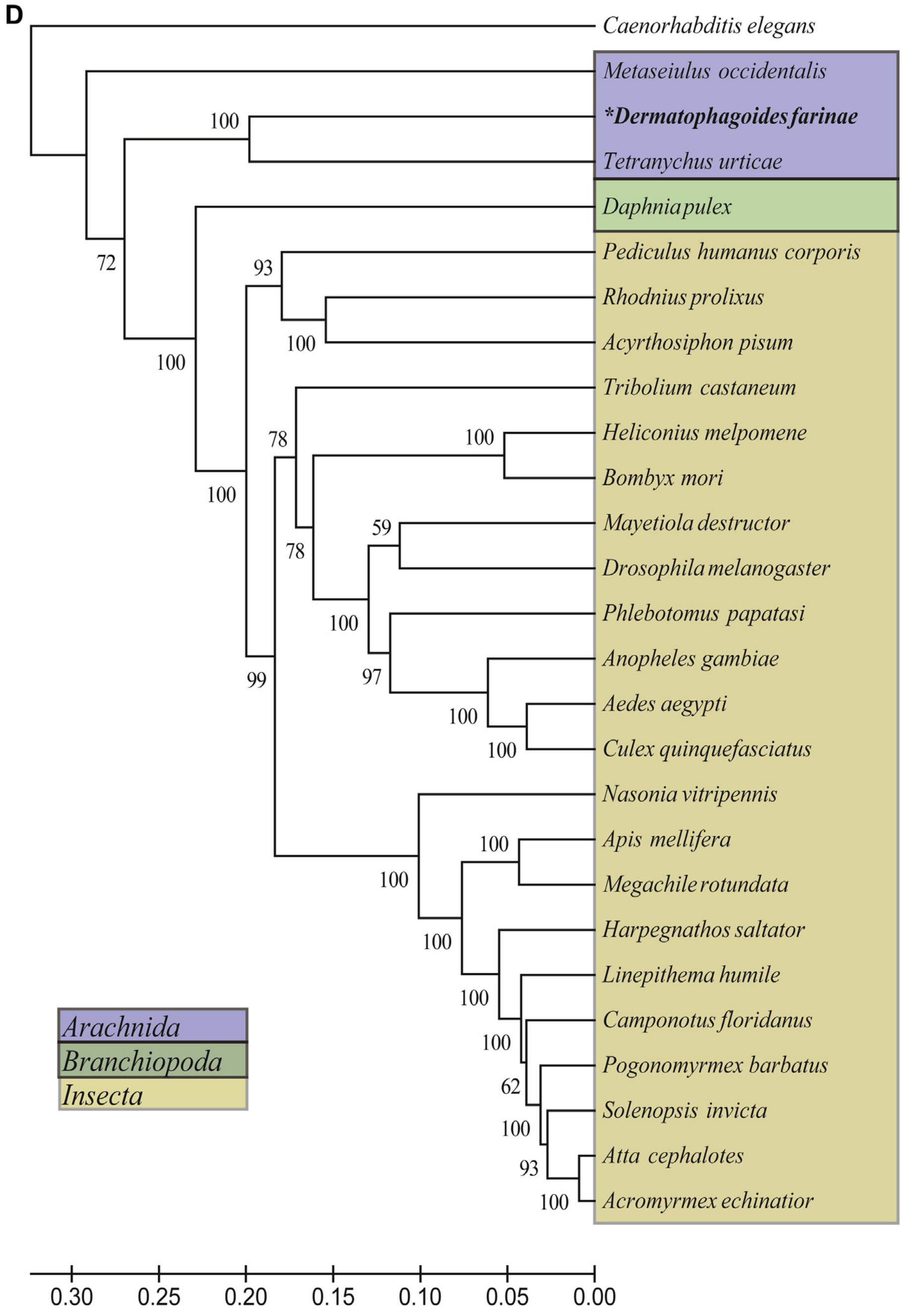

FIG 1. *(Continued).*

A rhodopsin–like G protein–coupled receptor, GPRadr2 (see Fig E5, *B*).

Identification of HDM allergen genes

We retrieved complete gene sequences and structures of 20 reported HDM allergens,[8,9] including Der f 1 to Der f 23, except for Der f 17, Group 12, and Group 19, from our assembled genome with support from transcriptomic analysis (Table I and see Fig E6 in this article's Online Repository at www.jacionline.org) and determined their relative expression levels in adult *D farinae* mites in terms of fragments per kilobase of transcript per million mapped reads (Table I). Der f 4 is reported here for the first time. We also identified the complete sequences of 7 noncanonical allergen candidates (profilin, Alt a 6, α-tubulin 1, cathepsin, Mala s 6, aldehyde dehydrogenase, and enolase homologues) with amino acid sequence homology (approximately 41.5% to 90.5%) to experimentally validated allergens in other species (see Fig E7 and Tables E8 and E9 in this article's Online Repository at www.jacionline.org). The structures of these 20 cloned canonical and 7 noncanonical allergen genes were mapped (see Figs E6 and E7) and their sequences were confirmed to be identical to sequences in our assembled *D farinae* genome, confirming they were of mite, rather than microbial, origin.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis of protein spots reactive to

TABLE I. Gene structures of canonical *D farinae* allergens confirmed by means of Sanger sequencing

Gene	Locus tag*	Biochemical function	No. of exons	Deduced no. of amino acids	FPKM	Homologue† (% similarity)
Der f 1	DEFA_073880	Cysteine protease	6	321	840.51	BAC53948 (100%)
Der f 2	DEFA_057430	Lipid binding	2	146	97.27	Q00855 (100%)
Der f 3	DEFA_036500	Trypsin	2	259	36.79	P49275 (99%)
Der f 4	DEFA_092370	α-Amylase	3	526	12.99	AAD38942 (88%)‡
Der f 5	DEFA_009370	Structural protein	2	132	410.29	ABO84970 (100%)
Der f 6	DEFA_160240	Chymotrypsin	3	279	24.83	ABG23667 (100%)
Der f 7	DEFA_012670	Unknown	2	213	481.87	ACK76298 (99%)
Der f 8	DEFA_112610	Glutathione transferase	2	221	353.39	AAP35080 (96%)‡
Der f 9	DEFA_108510	Serine protease	4	272	4.37	AAP57077 (92%)
Der f 10	DEFA_012620	Tropomyosin	5	284	1532.23	Q23939 (99%)
Der f 11	DEFA_029610	Paramyosin	11	876	321.77	AAO73464 (98%)
Der f 13	DEFA_016640	Fatty acid binding	2	131	1772.53	2A0A_A (100%)
Der f 14	DEFA_023480	Vitellogenin: egg yolk storage	6	1666	130.84	AAM21322 (88%)‡
Der f 15	DEFA_127470	Chitinase	4	556	40.83	AAD52672 (96%)
Der f 16	DEFA_053360	Gelsolin: actin binding	7	480	147.91	AAM64112 (99%)
Der f 18	DEFA_042810	Chitinase	3	462	88.28	AAM19082 (100%)
Der f 20	DEFA_122350	Arginine kinase	5	356	342.86	AAP57094 (99%)
Der f 21	DEFA_009360	Structural protein	2	136	943.53	AAX34048 (100%)
Der f 22	DEFA_072800	MD-2–related lipid recognition	2	155	820.98	ABG35122 (100%)
Der f 23	DEFA_123860	Chitin-binding domain type 2	3	91	31.67	ACB46292.1 (84%)

FPKM, Fragments per kilobase of transcript per million mapped reads.
*Locus tags are included in our assembled *D farinae* genome.
†Analyzed relative to the NCBI database (by BLAST algorithm); GenBank accession numbers are listed.
‡Partial.

pooled sera from 20 patients with HDM allergy with our integrated-omics approach (circled spots in Fig 2, *A* and *B*, and see Fig E8, *A* and *B*, in this article's Online Repository at www.jacionline.org) revealed 4 known canonical allergens (Der f 1, Der f 2, Der f 11, and Der f 14) and 12 other proteins (see Table E10 in this article's Online Repository at www.jacionline.org). The gene sequences of these 12 homologues were confirmed by using the Sanger method to be identical to sequences within our assembled *D farinae* genome. These sequencing data confirmed that these genes were from the mite genome.

On the basis of protein size and signal intensity, we selected 6 of these 12 proteins for production of recombinant proteins for probing of allergenicity: ubiquinol-cytochrome c reductase binding protein (UQCRB)–like protein; myosin alkali light-chain protein; secreted inorganic pyrophosphatase; DFP2; cofilin; and ferritin heavy-chain-like protein. Only recombinant UQCRB-like protein (see Fig E9 in this article's Online Repository at www.jacionline.org) was strongly bound by IgE in separate serum samples from 18 of 18 patients with HDM allergy and not bound by IgE antibodies in sera obtained from any of 18 nonallergic healthy control subjects and 7 patients with pollen allergy (Fig 2, *E*, and see Fig E10 and Table E11 in this article's Online Repository at www.jacionline.org). The other 2 recombinant proteins (cofilin and secreted inorganic pyrophosphatase) yielded very weak positive responses, whereas the remaining 3 recombinant proteins did not show any IgE-binding activities (data not shown). The gene encoding the UQCRB-like protein was cloned, and its gene structure was mapped (see Fig E11, *A*, in this article's Online Repository at www.jacionline.org). An IgE-ELISA demonstrated strong IgE binding to recombinant UQCRB-like protein ($Z = 5.6702$, $P < .0001$) in 22 (100%) of 22 sera from patients with mite allergy (Fig 2, *F*). *In vivo* testing showed skin reactivity in 5 of 10 patients with mite allergy (see Table E12 in this article's Online Repository at www.jacionline.org). This novel major allergen was designated as Der f 24 by the World Health Organization/International Union of Immunological Societies Allergen Nomenclature Sub-committee (http://www.allergen.org/viewallergen.php?aid=772).

Microbiome

Of the approximately 112,000 genomic sequencing reads that we linked to 100 microbial species (Fig 3, *A*, and see Table E5), 71,000 (63.4%) mapped uniquely to *Enterobacter* species, most predominantly to *E cloacae* and *Enterobacter hormaechei*; *Staphylococcus* (17.8%) and *Escherichia* (4.9%) species were the next most predominant genera (Fig 3, *A*, and see Table E5). *Bartonella* species accounted for only 1.7% of the reads. Our immunohistochemistry experiment confirmed the abundant presence of enterobacteria in the guts of *D farinae* (Fig 3, *B*).

We examined whether the genes encoding KEGG pathway enzymes were present in the *D farinae* genome or its microbiome fraction and identified 3 pathways wherein the majority of genes were from the microbiome: thiamine biosynthesis (see Fig E12 in this article's Online Repository at www.jacionline.org) and aromatic and aliphatic amino acid biosynthesis (see Figs E13 and E14 in this article's Online Repository at www.jacionline.org). Additionally, by combining the DNA and RNA sequence information, we determined from our draft genome that an intron is present in the 13.8-kDa *D farinae* bacteriolytic enzyme (Fig 3, *C*).

DISCUSSION

We produced a 53.5-Mb *D farinae* draft genome with 516 scaffolds and a complete 14.3-kb mitochondrial genome. Its

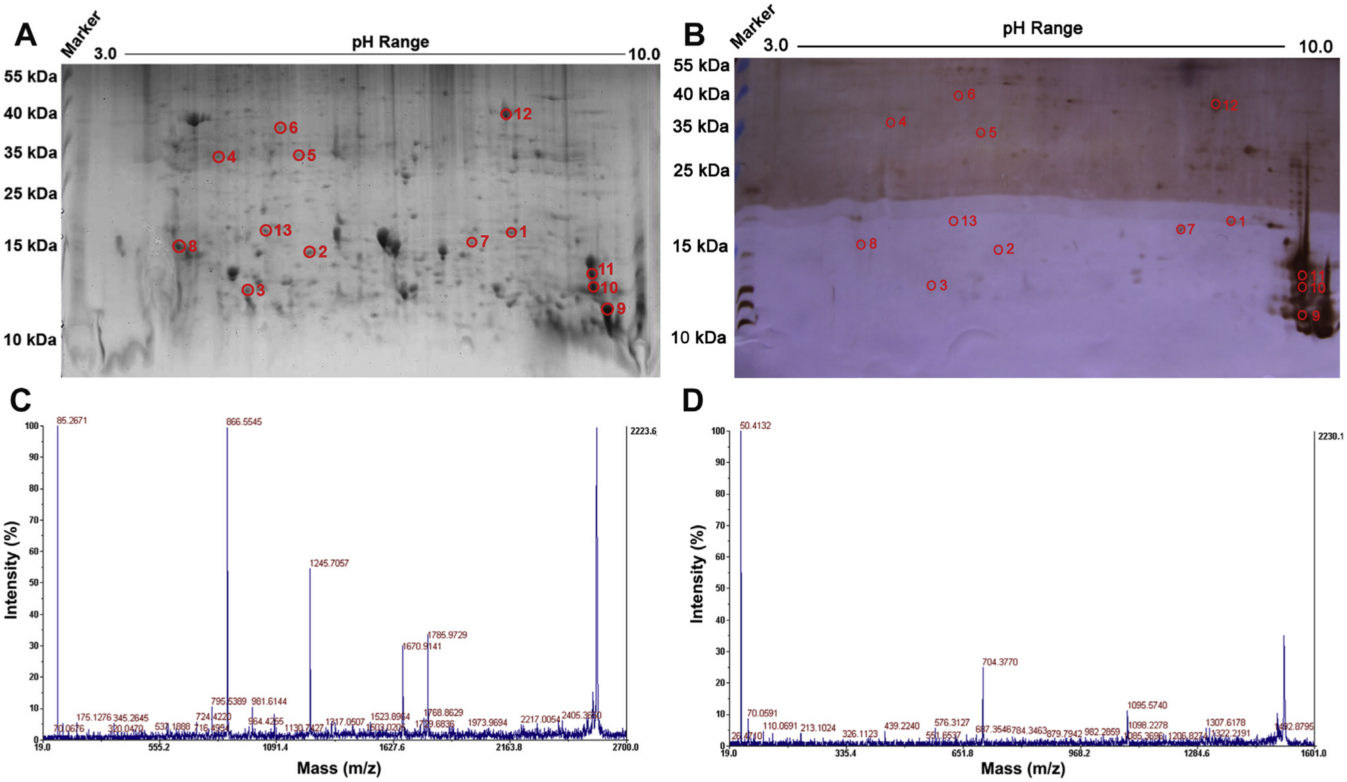

FIG 2. Proteomics discovery of UQCRB-like protein and IgE-binding assays. **A,** Two-dimensional polyacrylamide gel electrophoresis (isoelectric point range, 3.0-10.0) demonstrating 13 IgE-immunopositive spots *(red circles)*. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis results for each spot can be found in Table E10. **B,** Corresponding IgE-immunoblot image. **C,** Tandem mass spectrometry (MS/MS) spectrum of spot no. 3 matching peptide fragment YGLYYDDFYDYTDAAHLEAVR of the mite UQCRB-like protein; charts show intensity (relative abundance) of motifs as a function of m/z. **D,** Another MS/MS spectrum of spot no. 3 identical to peptide fragment LPPDLYDQHTYR of the mite UQCRB-like protein. **E,** Deduced amino acid sequence of the mite UQCRB-like protein (GenBank accession no.: KC669700). *Boxes* indicate the 2 fragments in MS/MS spectrums. **F,** Binding of recombinant UQCRB-like protein by IgE in sera from 18 of 18 patients with HDM allergy. *MW,* Molecular weight. **G,** ELISA demonstrating IgE binding to recombinant UQCRB-like protein (Z = 5.6702, *P* < .0001) in 22 (100%) of 22 sera from patients with HMD sensitization *(red circles)* and 22 nonallergic subjects *(blue squares)* as controls.

completeness was assessed by means of application of CEGMA to identify the existence of 248 CEGs that are present in a wide range of taxa.[26] As shown in Table E6, *A* and *B*, the number of complete proteins encoded was similar to that of *T urticae*,[13] with both reaching more than 95% completeness. Notably, we identified 3 additional complete proteins in our draft genome (see Table E6).

A draft genome provides a framework for an organism's DNA fragment assembly and allows up to 95% of genes to be identified. However, compared with a complete genome, it does have limitations. The first limitation is the potential omission of repetitive sequences in some introns or intergenic regions. Second, there might be a lack of correlation among scaffolds. However, these limitations will not affect most researchers who plan to do functional studies of *D farinae* genes. The draft genome can be further verified and filled out by means of physical mapping with recently developed optical mapping and mate-pair sequencing techniques,[29] as well as gap filling with traditional Sanger sequencing.

We confirmed that our *D farinae* genome draft includes 20 reported allergen genes from Der f 1 to Der f 23, except for Der f 17, Group 12, and Group 19, and verified 7 noncanonical allergen homologue genes within it. Der f 17 was reported by Tategaki et al in the Allergen Nomenclature Web page in 2000.[30] However, the sequence for Der f 17 has not been published. Thus far, no studies have reported Group 12 and Group 19 allergens in *D pteronyssinus* or *D farinae*. Using Blo t 12 and Blo t 19 as reference sequences, we performed BLAST searches in both the transcriptome and genome of *D farinae*, and no significant hits were identified. The lack of a hit could be due to the nonexistence of Group 12 and 19 allergens in *D farinae* or the incompleteness of our assembled draft *D farinae* genome (see Table E6, *A*). In addition, all 20 of the *D farinae* allergens described by An et al[31] were also confirmed to be located in our draft genome, and their full gene sequences and structures were recovered (see Table E13).

Furthermore, we found that the draft genome encoded a novel allergen, Der f 24, which is a UQCRB-like protein homologue that produced a strong reaction in all patients with HDM allergy who were tested. Our BLAST analysis and segregation of microbial DNA revealed *Enterobacter* species as the predominant bacterial genus inhabiting the mite. Our phylogenetic analysis revealed a strong correspondence between *D farinae* and *T urticae* genes encoding for metabolic processes. The divergence we observed (21.8% in *D farinae* and 19.1% in *T urticae* without a cross-species match) is likely

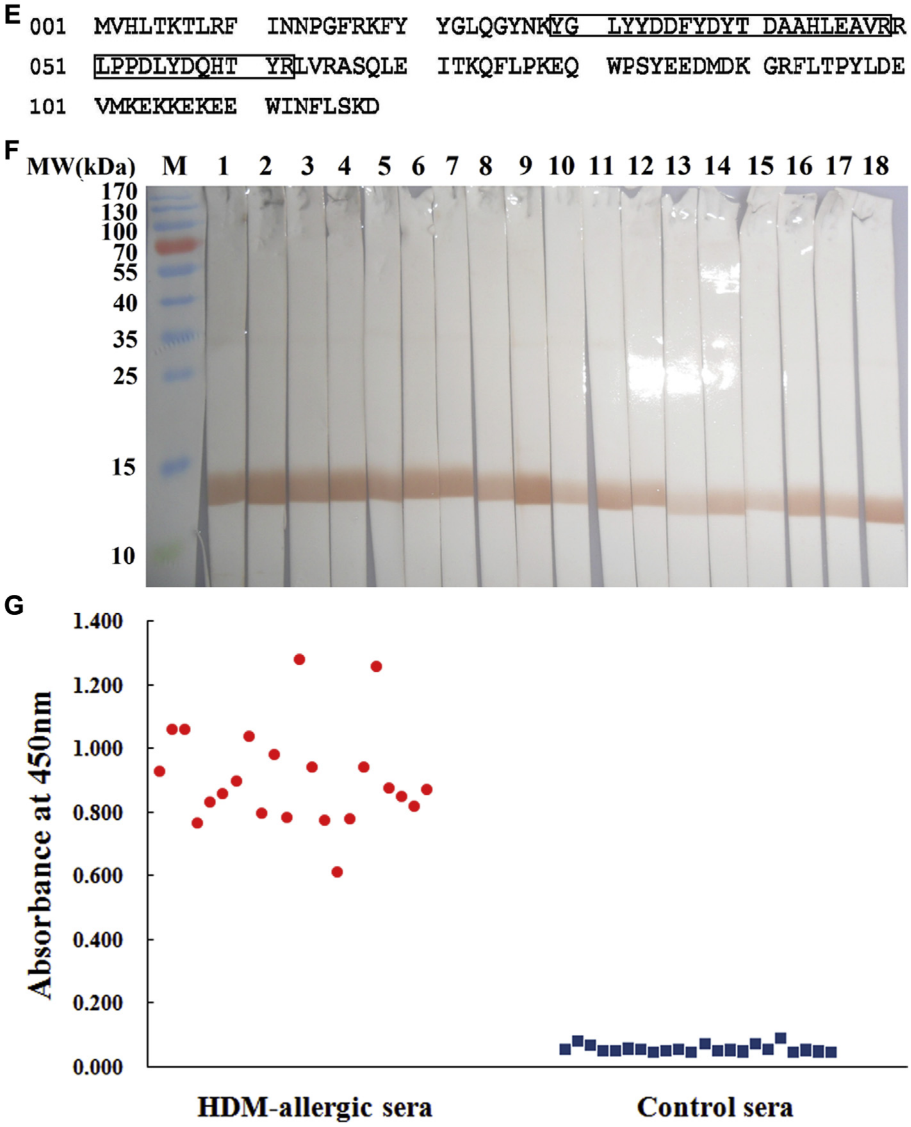

FIG 2. *(Continued).*

related to the species' differing living conditions. Our draft genome included annotated genes for a full phototransduction pathway, excepting rhodopsin, as well as a candidate gene (DEFA_098690) encoding a class A rhodopsin-like GPRadr2. Recovery of the GPRadr2 protein completed the phototransduction pathway, providing support for Furumizo's assertion that *D farinae* might have photoreceptors responsive to light in the 500- to 575-nm range.[32] Further investigation might provide molecular targets for the development of more effective acaricides to control HDM propagation.[33,34]

Elucidating the full spectrum of allergens from *D farinae* and *D pteronyssinus*, the most prevalent mite species, including determination of their biochemical function, gene structures, complete cDNA sequences, localization, content, epitope, IgE-binding activity, and skin test activity, is important for obtaining a good understanding of mite allergies.[5,6] The present *D farinae* draft genome and transcriptome could propel research in the HDM allergy field forward by serving as an essential resource for proteomic identification of novel allergens and enabling efficient and reliable identification of full-length allergen genes. Here, by using the *D farinae* transcriptome database, we were able to identify a novel allergen, namely Der f 24 (Fig 2). The observed molecular weight of Der f 1 in our 2-dimensional gel (spot no. 5 in Fig 2, *A*) was about 35 kDa. We expect that it is the proform or preproform of Der f 1 rather than the mature 25-kDa form. The IgE reactivity of pro–Der f 1 would be expected to be lower than that of mature Der f 1, and mature Der f 1 is hardly refolded after heating or denaturation.[35] Our observed molecular weights for Der f 11 and Der f 14 in 2-dimensional gels were much less than the theoretic weights (see Table E10), perhaps because of degradation during protein extraction. Except for Der f 1, Der f 2, Der f 11, and Der f 14, other canonical allergens were not identified in these proteomic experiments. The plausible contributory reasons are discussed in the Methods section in this article's Online Repository. In addition, because of inherent technological limitations (eg, minimum mass requisite), 68 of 86 IgE-reactive spots in the 2-dimensional gels could not be analyzed by means of mass spectrometry in our

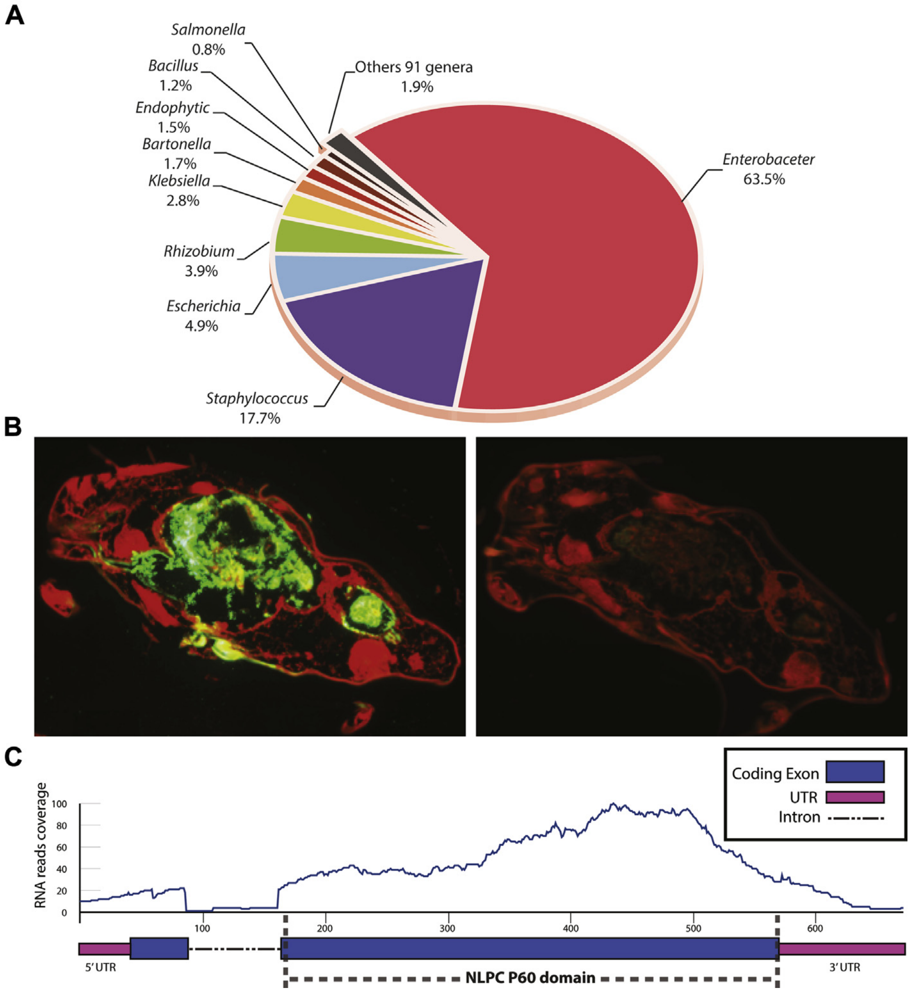

FIG 3. Endosymbiotic microbes and bacteriolytic enzyme. **A,** Distribution of bacterial genera living in *D farinae*. *Endophytic*, Unclassified bacteria. **B,** Distinctive labeling of gut contents with anti–*E cloacae* antibody *(left)* and no-antibody controls *(right)*. **C,** Structure of the *D farinae* gene (DEFA_122470) encoding 13.8-kDa bacteriolytic enzyme (2 exons and 1 intron). *UTR*, Untranslated region.

study (Fig 2, *B*, and see Fig E8, *A*). These uncircled IgE-reactive spots might be worth investigating in the future (Fig 2, *B*).

This is the first report of UQCRB-like protein as a major allergen. No protein with a similar biochemical function has been implicated as an allergen; UQCRB-like protein likely represents a new major allergen class. In our phylogenetic analysis (see Fig E11, *B*) *D farinae* UQCRB-like protein clustered with proteins of other arthropods but branched away from the cluster, underscoring its uniqueness.

Here we confirmed that the internal HDM body is host to more than 100 bacterial species (see Table E4). The dominant presence of *Enterobacter* species in the *D farinae* microbiome, rather than *Bartonella* species, as previously suggested, is noteworthy given their potential clinical importance; enterobacteria are isolated in approximately 10% of nosocomial respiratory tract infections, with 60% to 70% of those being *E cloacae*.[36-39] The HDMs might serve as an intermediate host for enterobacteria and contribute to their transmission. Additionally, it should be noted that the distribution of sequence reads obtained in this study might be biased toward mite genomic DNA rather than bacterial DNA. The preparation of the mite genomic library was not optimized for bacterial genomic DNA isolation, which could require extended protease, SDS, and lysozyme treatments, particularly for gram-positive species. Both the prior and present studies have reported 24 groups of dust mite allergens not of microbial origin. Our 2-dimensional polyacrylamide gel electrophoresis immunoblot experiment revealed 18 IgE-binding spots originating from *D farinae* but not its microbiota. Nembrini et al[40] found that airborne microbial products could suppress allergic inflammation through a multicomponent immunoregulatory mechanism. This finding might explain, at least in part, why bacterial proteins in dust mite bodies are not allergenic.

Previously, Mathaba et al[41] isolated a 13.8-kDa bacteriolytic enzyme from *D pteronyssinus* mite extracts and suggested it was derived from bacteria. Erban et al[42] cloned a homologous cDNA in *D farinae* mites with an oligo-dT primer, which implied that the bacteriolytic enzyme was of mite origin. Our data indicate that the gene that encodes this enzyme, DEFA_122470, contains an intron, which should not be in bacterial genes (Fig 3, *C*), supporting a mite origin. Such gut enzymes might lyse bacteria, enabling mites to obtain nutrition from them.[43] The mites died when exposed to ampicillin in culture (data not shown), indicating that they depend on their microbiome. Given the mite's reliance on bacteria for some essential nutrients, such as thiamine and aromatic amino acids (see Figs E12-E14), our observations support the view that there is a symbiotic relationship between *D farinae* and its gut microbes.

From our isolated microbial scaffolds, we identified 20 of the 30 genes known to encode enzymes involved in LPS biosynthesis (see Fig E15 in this article's Online Repository at www.jacionline.org). The presence of these genes in the *D farinae* microbiome supports the inclusion of bacterial endotoxins in mite allergen vaccines.[11] House dust endotoxin levels have been associated with increased asthma severity.[44] Conversely, because endotoxin is a potent inducer of T_H1-type cytokines, early indoor endotoxin exposure might protect infants against allergen sensitization.[45] Given the abundance of microbial life in the *D farinae* gut and the observation that *D farinae*'s digestive system constitutes approximately 70% of its body size,[46] we believe it will be important to take the HDM's microbiome into consideration to elucidate HDM biology, allergenicity, and immunotherapy mechanism.

In conclusion, a *D farinae* genome draft that revealed the full gene structures of 20 canonical mite allergens and 7 noncanonical allergen homologues was produced. The metagenomic landscape of the mite shown in our results revealed *Enterobacter* species as the predominant genus of the *D farinae* gut microbiome. The genome draft enabled us to pursue a multi-omic approach in which we identified Der f 24, a UQCRB homologue, as a novel major allergen. This draft genome provides a tool for future identification and characterization of HDM allergens and will be of use to developers of diagnostics and immunotherapeutic vaccines.

Key messages

- **A *D farinae* genome draft was produced and revealed the full gene structures of 20 canonical mite allergens and 7 noncanonical allergen homologues.**
- ***Enterobacter* species is the predominant genus in the *D farinae* gut microbiome.**
- **The genome draft enabled efficient identification of a UQCRB homologue, a novel major allergen designated Der f 24, through a multi-omic approach.**

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. Liu Z, Bai Y, Ji K, Liu X, Cai C, Yu H, et al. Detection of *Dermatophagoides farinae* in the dust of air conditioning filters. Int Arch Allergy Immunol 2007; 144:85-90.
3. 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;39:325-39.
4. 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.
5. 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.
6. Colloff M. Dust mites. Dordrecht (The Netherlands): CSIRO Publishing and Springer Science; 2009:273-328.
7. Weghofer M, Grote M, Resch Y, Casset A, Kneidinger M, Kopec J, et al. Identification of Der p 23, a peritrophin-like protein, as a new major *Dermatophagoides pteronyssinus* allergen associated with the peritrophic matrix of mite fecal pellets. J Immunol 2013;190:3059-67.
8. Shakib F, Ghaemmaghami AM, Sewell HF. The molecular basis of allergenicity. Trends Immunol 2008;29:633-42.
9. Trompette A, Divanovic S, Visintin A, Blanchard C, Hegde RS, Madan R, et al. Allergenicity resulting from functional mimicry of a Toll-like receptor complex protein. Nature 2009;457:585-8.
10. Pecoud A, Nicod L, Badan M, Agrell B, Dreborg S, Kolly M. Effects of one-year hyposensitization in allergic rhinitis. Comparison of two house dust mite extracts. Allergy 1990;45:386-92.
11. Trivedi B, Valerio C, Slater JE. Endotoxin content of standardized allergen vaccines. J Allergy Clin Immunol 2003;111:777-83.
12. Chen KW, Blatt K, Thomas WR, Swoboda I, Valent P, Valenta R, et al. Hypoallergenic Der p 1/Der p 2 combination vaccines for immunotherapy of house dust mite allergy. J Allergy Clin Immunol 2012;130:435-43.e4.
13. Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, Grbic V, et al. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 2011;479:487-92.
14. Delgado J, Orta JC, Navarro AM, Conde J, Martínez A, Martínez J, et al. Occupational allergy in greenhouse workers: sensitization to *Tetranychus urticae*. Clin Exp Allergy 1997;27:640-5.
15. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20: 265-72.
16. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011;108:1513-8.
17. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821-9.
18. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 2008;18:1979-90.
19. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 2004;20:2878-9.
20. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;25:955-64.
21. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35:3100-8.
22. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644-52.
23. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105-11.
24. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res 2010;38:4570-8.
25. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010;28:511-5.
26. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;23:1061-7.
27. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40: D109-14.
28. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674-6.
29. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010; 11:31-46.
30. Tategaki A, Kawamoto S, Aki T, Jyo T, Suzuki O, Shigeta S, et al. Newly described house dust mite allergens. ACI International 2011;1(suppl):74-6.
31. An S, Chen L, Long C, Liu X, Xu X, Lu X, et al. *Dermatophagoides farinae* allergens diversity identification by proteomics. Mol Cell Proteomics 2013;12: 1818-28.
32. Colloff M. Dust mites. Dordrecht (The Netherlands): CSIRO Publishing and Springer Science; 2009:84.
33. Arlian LG, Platts-Mills TA. The biology of dust mites and the remediation of mite allergens in allergic disease. J Allergy Clin Immunol 2001;107(suppl):S406-13.
34. Wu HQ, Li L, Li J, He ZD, Liu ZG, Zeng QQ, et al. Acaricidal activity of DHEMH, derived from patchouli oil, against house dust mite, *Dermatophagoides farinae*. Chem Pharm Bull (Tokyo) 2012;60:178-82.
35. Takai T, Kato T, Yasueda H, Okumura K, Ogawa H. Analysis of the structure and allergenicity of recombinant pro- and mature Der p 1 and Der f 1: major

conformational IgE epitopes blocked by prodomains. J Allergy Clin Immunol 2005;115:555-63.

36. Gastmeier P, Sohr D, Geffers C, Ruden H, Vonberg RP, Welte T. Early- and late-onset pneumonia: is this still a useful classification? Antimicrob Agents Chemother 2009;53:2714-8.

37. Sanders WE Jr, Sanders CC. *Enterobacter* spp.: pathogens poised to flourish at the turn of the century. Clin Microbiol Rev 1997;10:220-41.

38. Valerio CR, Murray P, Arlian LG, Slater JE. Bacterial 16S ribosomal DNA in house dust mite cultures. J Allergy Clin Immunol 2005;116:1296-300.

39. Millner PD, Ericson KE, Marsh PB. Bacteria on closed-boll and commercially harvested cotton. Appl Environ Microbiol 1982;44:355-62.

40. Nembrini C, Sichelstiel A, Kisielow J, Kurrer M, Kopf M, Marsland BJ. Bacterial-induced protection against allergic inflammation through a multi-component immunoregulatory mechanism. Thorax 2011;66:755-63.

41. Mathaba LT, Pope CH, Lenzo J, Hartofillis M, Peake H, Moritz RL, et al. Isolation and characterisation of a 13.8-kDa bacteriolytic enzyme from house dust mite extracts: homology with prokaryotic proteins suggests that the enzyme could be bacterially derived. FEMS Immunol Med Microbiol 2002;33:77-88.

42. Erban T, Di Presa CA, Kopecky J, Poltronieri P, Hubert J. PCR detection of the 14.5 antibacterial NlpC/P60-like *Dermatophagoides pteronyssinus* protein in *Dermatophagoides farinae* (Acari: Pyroglyphidae). J Med Entomol 2013;50:931-3.

43. Erban T, Hubert J. Digestive function of lysozyme in synanthropic acaridid mites enables utilization of bacteria as a food source. Exp Appl Acarol 2008;44:199-212.

44. Rizzo MC, Naspitz CK, Fernandez-Caldas E, Lockey RF, Mimica I, Sole D. Endotoxin exposure and symptoms in asthmatic children. Pediatr Allergy Immunol 1997;8:121-6.

45. Gereda JE, Leung DY, Thatayatikom A, Streib JE, Price MR, Klinnert MD, et al. Relation between house-dust endotoxin exposure, type 1 T-cell development, and allergen sensitisation in infants at high risk of asthma. Lancet 2000;355:1680-3.

46. Zhang YY, Sun X, Liu ZG. Morphology and three-dimensional reconstruction of the digestive system of *Dermatophagoides farinae*. Int Arch Allergy Immunol 2008;146:219-26.