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.