



**Genome Sequence of Aedes aegypti, a Major Arbovirus Vector**

Vishvanath Nene, *et al.*

*Science* **316**, 1718 (2007);

DOI: 10.1126/science.1138878

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of June 22, 2007):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/316/5832/1718>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/1138878/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/316/5832/1718#related-content>

This article **cites 30 articles**, 14 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/316/5832/1718#otherarticles>

This article has been **cited by 2** articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/316/5832/1718#otherarticles>

This article appears in the following **subject collections**:

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

# Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector

Vishvanath Nene,<sup>1\*</sup> Jennifer R. Wortman,<sup>1</sup> Daniel Lawson,<sup>2</sup> Brian Haas,<sup>1</sup> Chinnappa Kodira,<sup>3</sup> Zhijiang (Jake) Tu,<sup>4</sup> Brendan Loftus,<sup>1†</sup> Zhiyong Xi,<sup>5</sup> Karyn Megy,<sup>2</sup> Manfred Grabherr,<sup>3</sup> Qinghu Ren,<sup>1</sup> Evgeny M. Zdobnov,<sup>6,7,8</sup> Neil F. Lobo,<sup>9</sup> Kathryn S. Campbell,<sup>10</sup> Susan E. Brown,<sup>11</sup> Maria F. Bonaldo,<sup>12</sup> Jingsong Zhu,<sup>13</sup> Steven P. Sinkins,<sup>14</sup> David G. Hogenkamp,<sup>15‡</sup> Paolo Amedeo,<sup>1</sup> Peter Arensburger,<sup>13</sup> Peter W. Atkinson,<sup>13</sup> Shelby Bidwell,<sup>1</sup> Jim Biedler,<sup>4</sup> Ewan Birney,<sup>2</sup> Robert V. Bruggner,<sup>9</sup> Javier Costas,<sup>16</sup> Monique R. Coy,<sup>4</sup> Jonathan Crabtree,<sup>1</sup> Matt Crawford,<sup>3</sup> Becky deBruyn,<sup>9</sup> David DeCaprio,<sup>3</sup> Karin Eiglmeier,<sup>17</sup> Eric Eisenstadt,<sup>1</sup> Hamza El-Dorry,<sup>18</sup> William M. Gelbart,<sup>10</sup> Suely L. Gomes,<sup>18</sup> Martin Hammond,<sup>2</sup> Linda I. Hannick,<sup>1</sup> James R. Hogan,<sup>9</sup> Michael H. Holmes,<sup>1</sup> David Jaffe,<sup>3</sup> J. Spencer Johnston,<sup>19</sup> Ryan C. Kennedy,<sup>9</sup> Hean Koo,<sup>1</sup> Saul Kravitz,<sup>20</sup> Evgenia V. Kriventseva,<sup>6</sup> David Kulp,<sup>21</sup> Kurt LaButti,<sup>3</sup> Eduardo Lee,<sup>1</sup> Song Li,<sup>4</sup> Diane D. Lovin,<sup>9</sup> Chunhong Mao,<sup>4</sup> Evan Mauceli,<sup>3</sup> Carlos F. M. Menck,<sup>22</sup> Jason R. Miller,<sup>1</sup> Philip Montgomery,<sup>3</sup> Akio Mori,<sup>9</sup> Ana L. Nascimento,<sup>23</sup> Horacio F. Naveira,<sup>24</sup> Chad Nusbaum,<sup>3</sup> Sinéad O'Leary,<sup>3</sup> Joshua Orvis,<sup>1</sup> Mihaela Perlea,<sup>1§</sup> Hadi Quesneville,<sup>25</sup> Kyanne R. Reidenbach,<sup>15</sup> Yu-Hui Rogers,<sup>20</sup> Charles W. Roth,<sup>17</sup> Jennifer R. Schneider,<sup>9</sup> Michael Schatz,<sup>1§</sup> Martin Shumway,<sup>1</sup> Mario Stanke,<sup>26,27</sup> Eric O. Stinson,<sup>9</sup> Jose M. C. Tubio,<sup>28</sup> Janice P. VanZee,<sup>15</sup> Sergio Verjovski-Almeida,<sup>18</sup> Doreen Werner,<sup>27</sup> Owen White,<sup>1</sup> Stefan Wyder,<sup>6</sup> Qiandong Zeng,<sup>3</sup> Qi Zhao,<sup>1</sup> Yongmei Zhao,<sup>1</sup> Catherine A. Hill,<sup>15</sup> Alexander S. Raikhel,<sup>13</sup> Marcelo B. Soares,<sup>12</sup> Dennis L. Knudson,<sup>11</sup> Norman H. Lee,<sup>1||</sup> James Galagan,<sup>3</sup> Steven L. Salzberg,<sup>1§</sup> Ian T. Paulsen,<sup>1</sup> George Dimopoulos,<sup>5</sup> Frank H. Collins,<sup>9</sup> Bruce Birren,<sup>3</sup> Claire M. Fraser-Liggett,<sup>1#</sup> David W. Severson<sup>9\*</sup>

We present a draft sequence of the genome of *Aedes aegypti*, the primary vector for yellow fever and dengue fever, which at ~1376 million base pairs is about 5 times the size of the genome of the malaria vector *Anopheles gambiae*. Nearly 50% of the *Ae. aegypti* genome consists of transposable elements. These contribute to a factor of ~4 to 6 increase in average gene length and in sizes of intergenic regions relative to *An. gambiae* and *Drosophila melanogaster*. Nonetheless, chromosomal synteny is generally maintained among all three insects, although conservation of orthologous gene order is higher (by a factor of ~2) between the mosquito species than between either of them and the fruit fly. An increase in genes encoding odorant binding, cytochrome P450, and cuticle domains relative to *An. gambiae* suggests that members of these protein families underpin some of the biological differences between the two mosquito species.

Mosquitoes are vectors of many important human diseases. Transmission of arboviruses is largely associated with the subfamily Culicinae, lymphatic filarial worms with both the Culicinae and the subfamily Anophelinae, and transmission of malaria-causing parasites with the Anophelinae (1). *Aedes aegypti* is the best-characterized species within the Culicinae (2), primarily because of its easy transition from field to laboratory culture, and has provided much of the existing information on mosquito biology, physiology, genetics, and vector competence (3, 4). It maintains close association with human populations and is the principal vector of the etiological agents of yellow fever and dengue fever (5, 6), as well as for the recent chikungunya fever epidemics in countries in the Indian Ocean area (7). Despite an effective vaccine, yellow fever remains a disease burden in Africa and parts of South America, with ~200,000 cases per year resulting in ~30,000 deaths (5). About 2.5 billion people are at risk for dengue, with ~50 million cases per year and ~500,000 cases of dengue hemorrhagic fever,

the more serious manifestation of disease. The incidence of dengue, for which mosquito management is currently the only prevention option, is on the increase (8). Thus, there is an urgent need to improve the control of these diseases and their vector.

The availability of a draft sequence of the ~278 million base pair (Mbp) genome of *Anopheles gambiae* (9) has accelerated research to develop new mosquito- and malaria-control strategies. Comparisons between *An. gambiae* and *Drosophila melanogaster* (10) revealed genomic differences between the two insects that reflect their divergence ~250 million years ago (11). *Anopheles* mosquitoes radiated from the *Aedes* and *Culex* lineages ~150 million years ago (12), and *Ae. aegypti* and *An. gambiae* share similar characteristics such as anthropophily, but they exhibit variation in morphology and physiology, mating behavior, oviposition preferences, dispersal, and biting cycle (1). Both mosquito species have three pairs of chromosomes, but *Ae. aegypti* lacks heteromorphic sex chromosomes (13). To provide genomics platforms for

research into *Ae. aegypti* and to harness the power of comparative genome analyses, we undertook a project to sequence the genome of this mosquito species.

**Assembly of a draft genome sequence of *Aedes aegypti*.** Whole-genome shotgun sequencing was performed on DNA purified from newly hatched larvae of an inbred substrain (LVP<sup>ib12</sup>) of the Liverpool strain of *Ae. aegypti*, which is tolerant to inbreeding while maintaining relevant phenotypes (14). About 98% of the sequence, assembled using Arachne (15), is contained within 1257 scaffolds with an N50 scaffold size of ~1.5 Mbp (i.e., half of the assembly resides in scaffolds this size or longer). Assembly statistics for the 1376-Mbp genome are given in table S1. Data related to the genome project have been deposited in GenBank (project accession number AAGE00000000).

The genome size of *Ae. aegypti* as determined by sequence analysis is larger than the

<sup>1</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. <sup>2</sup>European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. <sup>3</sup>Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02141, USA. <sup>4</sup>Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. <sup>5</sup>Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA. <sup>6</sup>University of Geneva Medical School, 1 rue Michel-Servet, Geneva 1211, Switzerland. <sup>7</sup>Swiss Institute of Bioinformatics, 1 rue Michel-Servet, Geneva 1211, Switzerland. <sup>8</sup>Imperial College London, South Kensington Campus, London SW7 2AZ, UK. <sup>9</sup>University of Notre Dame, Notre Dame, IN 46556, USA. <sup>10</sup>Harvard University, Cambridge, MA 02138, USA. <sup>11</sup>College of Agricultural Sciences, Colorado State University, Fort Collins, CO 80523, USA. <sup>12</sup>Northwestern University, Chicago, IL 60614, USA. <sup>13</sup>University of California, Riverside, CA 92521, USA. <sup>14</sup>University of Oxford, Oxford OX1 3PS, UK. <sup>15</sup>Purdue University, West Lafayette, IN 47907, USA. <sup>16</sup>Centro Nacional de Genotipado, Fundación Pública Galega de Medicina Xenómica, Hospital Clínico Universitario de Santiago, Edif. Consultas Planta-2, Santiago de Compostela E-15706, Spain. <sup>17</sup>Institut Pasteur, Paris 75724, France. <sup>18</sup>Universidade de Sao Paulo, Instituto de Química, Sao Paulo SP 05508-900, Brazil. <sup>19</sup>Texas A&M University, College Station, TX 77843, USA. <sup>20</sup>Joint Technology Center, 5 Research Place, Rockville, MD 20850, USA. <sup>21</sup>University of Massachusetts, Amherst, MA 01003, USA. <sup>22</sup>Universidade de Sao Paulo, Instituto de Biomedical Sciences, Sao Paulo SP 05508-900, Brazil. <sup>23</sup>Instituto Butantan, Sao Paulo SP 05503-900, Brazil. <sup>24</sup>Universidade da Coruña, A Coruña 15001, Spain. <sup>25</sup>Institut Jacques Monod, CNRS, Université Paris Diderot et Université Pierre-et-Marie Curie 2, Place Jussieu, Paris 75252, France. <sup>26</sup>507A Engineering 2, University of California, 1156 High Street, Santa Cruz, CA 95064, USA. <sup>27</sup>Universität Göttingen, Goldschmidtstraße 1, Göttingen 37077, Germany. <sup>28</sup>Complexo Hospitalario Universitario de Santiago, Santiago de Compostela 15706, Spain.

\*To whom correspondence should be addressed. E-mail: nene@tigr.org (V.N.); severson.1@nd.edu (D.W.S.)

†Present address: University College Dublin, Dublin 4, Ireland.

‡Deceased.

§Present address: 3125 Biomolecular Sciences Building, University of Maryland, College Park, MD 20742, USA.

||Present address: George Washington University Medical Center, Ross Hall, Room 603, 2300 I Street, NW, Washington, DC 20037, USA.

#Present address: Institute of Genome Sciences and Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

original estimate, ~813 Mbp, which was based on  $C_0t$  (DNA reassociation kinetics) analysis carried out in 1991 (16). An overinflated genome size could arise from assembled sequence data as a result of allelic sequence polymorphism present in a heterogeneous population of mosquitoes being sequenced. Although the estimate of 1376 Mbp may contain some such regions, we do not believe that our estimate is out of range by a large margin for the following reasons: (i) The strain that was used for the sequencing project was highly inbred (14); (ii) assembled sequences that are potentially “undercollapsed” are <5% of the estimated genome size (fig. S1); and (iii) flow cytometry data from six isolates of *Ae. aegypti*, including the parent of LVP<sup>ib12</sup>, indicate estimated genome sizes of 1213 to 1369 Mbp (table S2).

Genetic and physical mapping data allowed assignment, but without order or orientation, of 63, 48, 39, 43, and 45 scaffolds to *Ae. aegypti* chromosome 1 and chromosome arms 2p, 2q, 3p, and 3q, respectively (14). These scaffolds total ~430 Mbp in length and represent ~31% of the genome (table S3). Thus, the development of high-resolution physical mapping techniques and the generation of additional random or targeted sequence data are priorities for improving the quality of the current fragmented genome assembly and size estimate. Such progress would enable unambiguous differentiation between regions of segmental duplications and residual haplotype polymorphism.

**The genome of *Aedes aegypti* is riddled with transposable elements.** Transposable elements (TEs) contribute substantially to the factor of ~5 size difference between the *Ae. aegypti* and *An. gambiae* genomes. About 47% of the *Ae. aegypti* genome consists of TEs (Fig. 1 and table S4; see table S4 legend for definitions of TE family, element, and copy). *Aedes aegypti* harbors all known types of TEs that have been reported in *An. gambiae* with the exception of two DNA transposons, *merlin* (17) and *gambol*

(18). Simple and tandem repeats occupy ~6% of the genome, and an additional ~15% consists of repetitive sequences that remain to be classified.

Most eukaryotic TE families characterized to date (19) are present in *Ae. aegypti* and more than 1000 TEs have been annotated, representing a diverse collection of TEs in a single genome (table S4). Although the majority of protein-coding TEs appear to be degenerate, more than 200 elements have at least one copy with an intact open reading frame (ORF) and other features suggesting recent transposition. About 3% of the genome is composed of ~13,000 copies of the element *Juan-A* in the Jockey family of non-long terminal repeat (LTR) retrotransposons. A tRNA-related short interspersed nuclear element, *Feilai-B*, has the highest copy number, with ~50,000 copies per haploid genome. Only one highly degenerate *mariner* element is found in *Ae. aegypti*, whereas at least 20 *mariner* elements, many with intact ORFs, were found in *An. gambiae*. TEs present in *Ae. aegypti* but missing from *An. gambiae* include the LOA family of non-LTR retrotransposons, the *Osvaldo* element of the *Ty3/gypsy* LTR retrotransposons (20), and a unique family, *Penelope* (21). Comparison of *Ae. aegypti* and *An. gambiae* TE sequences is consistent with the interpretation of an overall lack of apparent horizontal transfer events, as a single candidate for such events was identified (14); one full-length copy of the *ITmD37E* DNA transposon in *Ae. aegypti* is 93% identical at the nucleotide level to a similarly classified TE in *An. gambiae*.

Miniature inverted repeat transposable elements (MITEs) and MITE-like elements of non-protein-coding TEs in *Ae. aegypti* have terminal inverted repeat sequences and target-site duplications, features characteristic of transposition of DNA transposons. Such TEs can be mobilized to transpose in trans, by transposases encoded by DNA transposons (22). The latter TEs occupy only 3% of the *Ae. aegypti* genome and

they are less numerous than non-protein-coding DNA elements, which occupy 16% of the genome (table S4). Thus, DNA transposons may have contributed to the expansion in size and organization of the *Ae. aegypti* genome through cross-mobilization of MITEs and MITE-like TEs.

#### Annotation of the draft genome sequence.

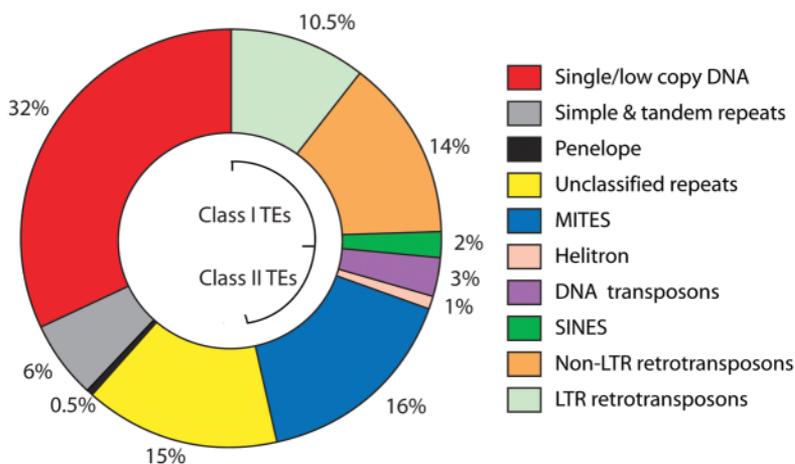
The fragmented nature of the assembled genome sequence, an asymmetric distribution of intron lengths within genes (figs. S2 and S3), and the frequent occurrence of TE-associated ORFs close to genes and within introns complicated the process of automated gene modeling and often led to prediction of split or chimeric gene models. Thus, we developed a multistage genome masking strategy to minimize the negative effects of TEs and other repetitive elements before gene finding (resulting in masking ~70% of the genome sequence). We also optimized gene-finding programs via iterative manual inspection of predicted gene models relative to a training set (14).

Two independent automated pipelines for structural annotation resulted in the prediction of 17,776 and 27,284 gene models, respectively (14). We made extensive use of a large collection of ~265,000 *Ae. aegypti* expressed sequence tags (ESTs) and dipteran protein and cDNA sequences in producing and then merging the two data sets into a single high-confidence gene set, which consists of 15,419 gene models (AaegL1.1). Alternative splice forms derived from these genes are predicted to generate at least 16,789 transcripts. Table 1 lists some of the genome and protein-coding characteristics of *Ae. aegypti* and those of *D. melanogaster* and *An. gambiae*.

Gene descriptions and molecular function Gene Ontology (GO) codes were assigned computationally to predicted protein sequences by means of BLASTP comparison searches with protein databases (14). The functional annotation pipeline included analyses of protein domains as well as secretion signal sequence and transmembrane motifs. A total of 8332 proteins were assigned a description, 9335 proteins were assigned GO terms, 2796 were assigned as “hypothetical proteins,” and 5027 were denoted “conserved hypothetical proteins.”

Genes encoding proteins <50 amino acid residues in length were not included in this annotation release unless they encoded known small proteins. However, these and other genes are captured in a set of 15,396 lower-confidence gene models that is available for analysis as a supplementary release (14). On the basis of transcriptional mapping data and limited manual examination, we anticipate that ~5 to 10% of the second-tier models or modified versions of them represent “real” genes.

**TEs contribute to complex protein-coding gene structures in *Aedes aegypti*.** A striking feature of protein-coding genes in *Ae. aegypti* is the factor of 4 to 6 increase in the average



**Fig. 1.** Relative genomic content of annotated TEs and other sequences in *Aedes aegypti*. TEs have been deposited in TEfam, a relational database for submission, retrieval, and analysis of TEs (<http://tefam.biochem.vt.edu>).

length of a gene relative to *An. gambiae* and *D. melanogaster*, which is due to longer intron lengths rather than longer exons or an increased number of introns (Table 1). The increased length of introns is primarily due to infiltration by TEs; a plot of intron size before and after masking repeat sequences reveals a shift to shorter intron lengths (fig. S2). A more global perspective of the genome expansion was revealed by the difference in genomic span (factor of ~4.6) of conserved gene arrangements between *Ae. aegypti* and *An. gambiae* that occupy ~33% of each genome (table S5 and fig. S3), providing evidence that TE-mediated expansion in both genic and intergenic regions has contributed to the increased size of the *Ae. aegypti* genome. Long introns, in particular those in 5' and 3' untranslated regions, are likely to complicate in silico studies to define cis-acting transcription and translational regulatory elements, as they may be distant from coding sequences (fig. S4).

**Transcriptional analyses.** Data derived from three different transcript-profiling platforms—ESTs, massively parallel signature sequencing (MPSS), and 60-nucleotide oligomer-based microarrays—were used to experimentally confirm predicted protein-coding gene models and to gain insight into differential transcription profiles (14). In total, the platforms identified transcripts from 12,350 (80%) of 15,419 genes. Mapping of ~265,000 ESTs and cDNA sequences and MPSS signature sequence tags to the genome sequence as well as gene models provided evidence for transcription of 9270 and 3984 genes, respectively, whereas microarray data identified transcripts from 9143 genes (table S6). The smaller number of genes identified by MPSS (table S7) may in part be explained by the observation that only about two-thirds of the genes can be assayed by MPSS, as this approach required the presence of a Dpn II restriction enzyme site within the transcribed region. The platforms identified a common set of 2558 genes and each platform identified a unique set of genes (fig. S5), which highlights the importance of using a multi-

platform approach. The data provide empirical support for ~76% of genes annotated as hypothetical (table S8), underscoring the validity of ab initio gene-finding programs in identifying novel genes.

Differences in transcript abundance between pools of RNA from nonadult developmental stages and from 4-day-old, non-blood-fed adult females were revealed by the microarray analyses, which identified 398 and 208 preadult stage and adult female enriched transcripts, respectively (table S9). Functional categorization of these transcripts differed mainly with regard to cytoskeletal, structural, and chemosensory functions (Fig. 2). Differential transcription of genes thought to be involved in chemosensory processes between these stages was conspicuous, with 17 transcripts highly enriched in mosquito developmental stages and only 3 enriched in adult females. A larger number of immune-system gene transcripts were also enriched in preadult stages (38 preadult versus 19 adult), which may reflect a broader microbial exposure of larvae and pupae in their aqueous environments. In addition, highly expressed genes encoding cuticle proteins in preadult stages (38 preadult versus 1 adult) are indicative of their function in cuticle metabolism and in a variety of other processes, including immunity, that are particularly dominant during development. The non-blood-fed status of the female mosquitoes did not enable discrimination of genes that carry out female-specific functions that mostly relate to blood processing and egg production.

***Aedes aegypti* gene families and domain composition.** Consistent with evolutionary distance estimates (12), there is a higher degree of similarity between the *Ae. aegypti* and *An. gambiae* proteomes than between the mosquito and *D. melanogaster* proteomes. Orthologous proteins were computed among the three genomes, with 67% of the *Ae. aegypti* proteins having an ortholog in *An. gambiae* and 58% having an ortholog in *D. melanogaster* (Fig. 3A). Analysis of three-way, single-copy orthologs revealed average amino acid identity of 74% between

the mosquito proteins, in contrast with ~58% identity between mosquito and fruit fly proteins (fig. S7). About 2000 orthologs are shared only between the mosquitoes and may represent functions central to mosquito biology. Although most of these proteins are of unknown function, ~250 can be assigned a predicted function, of which 28% are involved in gustatory or olfactory systems, 12% are members of the cuticular gene family, and 8% are members of the cytochrome P450 family (14).

Mapping of protein domains with Interpro (23) revealed an expansion of zinc fingers, insect cuticle, chitin-binding peritrophin-A, cytochrome P450, odorant binding protein (OBP) A10/OS-D, and insect allergen-related domains, among others, in *Ae. aegypti* relative to *An. gambiae*, *D. melanogaster*, and the honey bee *Apis mellifera* (table S10). Some of these constitute large *Ae. aegypti* gene families, as revealed by two independent clustering methods (14) (table S11). Genes containing zinc finger-like domains could be of transposon or retroviral origin, and these remain to be assessed.

Species-specific differences in the number of members within a multigene family often provide clues about biological adaptation to environmental challenges. In this context, cuticle proteins have been described to play diverse roles in exoskeleton formation and wound healing and are expressed in hemocytes, a major cell type that mediates innate immunity (24). Cuticular proteins also are implicated in arbovirus transmission (25). Expansion of olfactory receptors and OBPs in *Ae. aegypti* may contribute to an elaborate olfactory system, which in turn may be linked to the expansion in detoxification capacity. The latter and insect allergen-related genes, suggested to have a digestive function, may contribute to the relative robustness of *Ae. aegypti* and also could manifest in a higher insecticide resistance. In this context, the genome and EST data have led to the development of a specific microarray to identify candidate genes among members of multigene families (cytochrome P450, glutathione *S*-transferase, and carboxylesterase) associated with metabolic resistance to insecticides (26). This platform will provide a means to rapidly survey mechanisms of insecticide resistance in mosquito populations and represents an important tool in managing insecticide deployment and development programs.

G protein-coupled receptors (GPCRs) that are expected to function in signal transduction cascades in *Ae. aegypti* have been manually identified (14). This superfamily of proteins includes 111 nonsensory class A, B, and C GPCRs, 14 atypical class D GPCRs, and 10 opsin photoreceptors (tables S12 and S13). *Aedes aegypti* possesses orthologs for >85% of the *An. gambiae* and *D. melanogaster* nonsensory GPCRs, which suggests conservation of GPCR-mediated neurological processes across the Diptera. Many *Ae. aegypti* GPCRs have sequence similarity to known drug targets (27) and may reveal new

**Table 1.** Comparative statistics of *Ae. aegypti* nuclear genome coding characteristics.

Feature	Species		
	<i>Ae. aegypti</i>	<i>An. gambiae</i> †	<i>D. melanogaster</i> ‡
Size (Mbp)	1,376	272.9	118
Number of chromosomes	3	3	4
Total G+C composition (%)	38.2	40.9	42.5
Number of protein-coding genes	15,419	13,111	13,718
Average gene length* (bp)	14,587	5,124	3,460
Average protein-coding gene length† (bp)	1,397	1,154	1,693
Percent genes with introns	90.1	93.6	86.2
Average number of exons/gene	4.0	3.9	4.9
Average intron length (bp)	4,685	808	1,175
Longest intron (bp)	329,294	87,786	132,737
Average length of intergenic region (bp)	56,417	17,265	6,043

\*Includes introns but not untranslated regions. †Does not include introns. ‡Statistics were derived from genome updates for *An. gambiae* R-AgamP3 and *D. melanogaster* R-4.2.

opportunities for the development of novel insecticides.

**Metabolic potential and membrane transporters.** *Aedes aegypti* and *An. gambiae* are predicted to contain similar metabolic profiles as judged by assigning an Enzyme Commission (EC) number to both mosquito proteomes (table S14). Given the early stages of annotation, it is premature to draw conclusions from missing enzymes in predicted *Ae. aegypti* metabolic pathways. For example, assignment of EC numbers to the supplemental *Ae. aegypti* gene set (table S14) resulted in the identification of an additional 12 EC numbers (table S15) not present in AeGL1.1.

An automated pipeline (28) was used to predict potential membrane transporters for *Ae.*

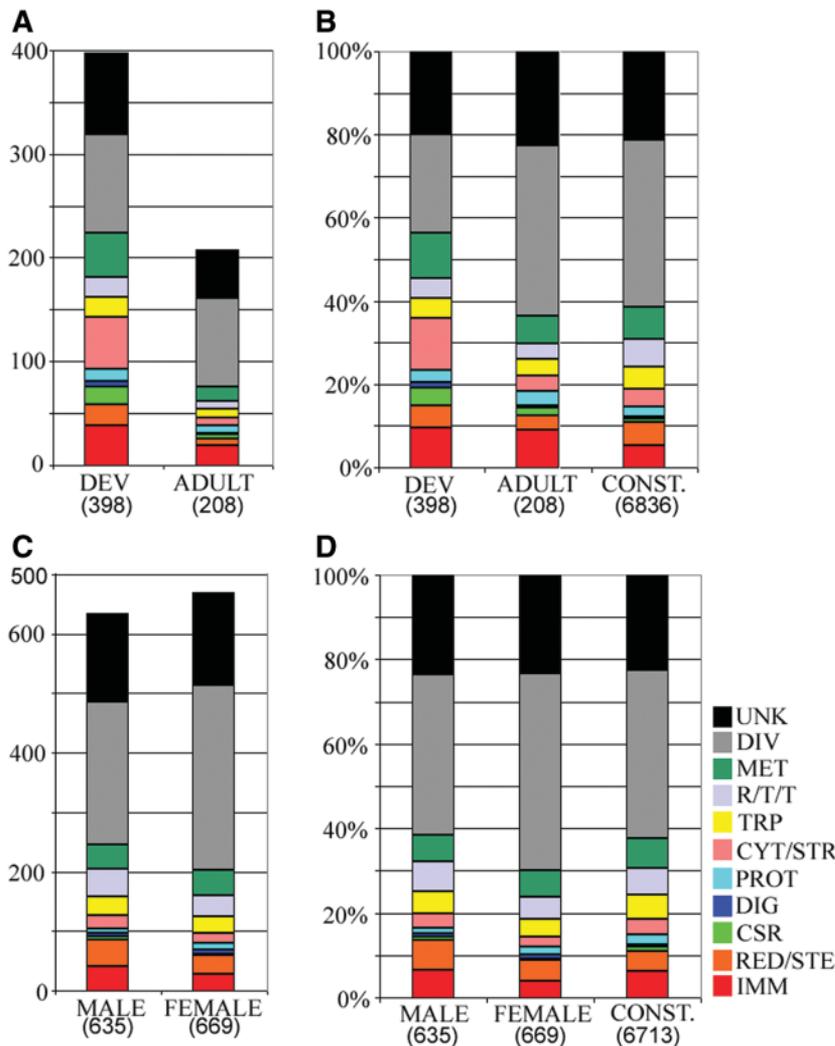
*aegypti* and *An. gambiae*, and their transport capacity resembles that of *D. melanogaster* (table S16). Similar to other multicellular eukaryotes, ~32% of all three insect transporters code for ion channels and probably function to maintain hemolymph homeostasis under different environmental conditions by modulating the concentrations of Na<sup>+</sup>, K<sup>+</sup>, and Cl<sup>-</sup> ions. *Aedes aegypti* encodes 52 more paralogs of voltage-gated potassium ion channels, epithelial sodium channels, and ligand-gated ion channels than *An. gambiae* and 65 more such paralogs than *D. melanogaster*. These channels play important roles in the signal transduction pathway and cell communication in the central nervous system and at neuromuscular junctions. A collection of

64 putative adenosine triphosphate-binding cassette transporters was identified, including subgroups that encode multidrug efflux proteins. *Aedes aegypti* encodes 16 more members of four different types of amino acid transporters than *An. gambiae* and 13 more members than *D. melanogaster*. Mosquito larvae cannot synthesize de novo all the basic, neutral, or aromatic L-amino acids (3) and must rely on uptake of these essential amino acids. The richer repertoire of membrane transport systems in *Ae. aegypti* is likely to intersect with the apparent increase in odorant reception and detoxification capacity.

**Autosomal sex determination and sex-specific gene expression.** Heteromorphic sex chromosomes are absent in *Ae. aegypti* and other culicine mosquitoes (13). Instead, sex is controlled by an autosomal locus wherein the male-determining allele, *M*, is dominant. The primary switch mechanism at the top of the mosquito sex determination cascade is different from that of *D. melanogaster*, where the X-chromosome/autosome ratio controls sex differentiation. However, we expect conservation of function in mosquito orthologs of *Drosophila* genes that are further downstream of the cascade (29). We verified the presence of a number of these in *Ae. aegypti*, including orthologs for *doublesex*, *transformer-2*, *fruitless*, *dissatisfaction*, and *intersex* (table S17).

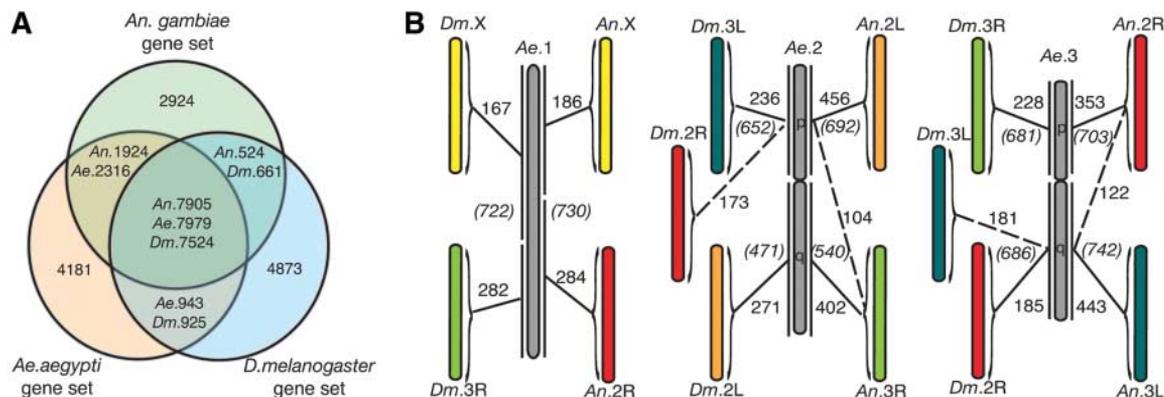
To define gene expression differences between the sexes, we analyzed microarray transcription profiles of 4-day-old, non-blood-fed adult female and male mosquitoes (Fig. 2); 669 and 635 transcripts were enriched in females and males, respectively, and 6713 transcripts were expressed at similar levels in both sexes (table S18). An additional 373 and 534 transcripts generated exclusive hybridization signals (with signal intensity below the cutoff threshold level in one channel) in females and males, respectively, and may therefore represent sex-specific transcripts. Functional categorization of female and male enriched transcripts yielded similar results, with some exceptions; male mosquitoes expressed a larger number of immune system-related transcripts (40 in males versus 25 in females) and redox- or stress-related transcripts (45 in males versus 33 in females). By comparing the *Ae. aegypti* profiles with previously described *An. gambiae* sex-specific microarray analyses (30), we identified 144 orthologous genes displaying the same sex-specific transcription pattern in *An. gambiae* (table S19), whereas 74 orthologs showed an opposite profile (table S20), suggesting differences in certain sex-specific functions between the two mosquito species.

**Conserved synteny with *Anopheles gambiae* and *Drosophila melanogaster*.** The assignment of 238 *Ae. aegypti* scaffolds containing ~5000 genes—about one-third of the predicted gene set—to a chromosomal location on the basis of genetic and physical mapping data (14) allowed us to compare ortholog position and to identify



**Fig. 2.** Transcriptome analyses of *Aedes aegypti*. (A) Functional class distributions of genes that are enriched in preadult stages (DEV) and the adult female stage (ADULT) (table S9). (B) Proportions of functional gene classes, expressed as percentage of the total number of genes that are enriched in preadult stages (DEV), adult female stage (ADULT), and constitutively expressed genes (CONST.). (C and D) same as (A) and (B) for genes enriched in the male, in the female, and common (CONST.) for both sexes (table S18). Functional classes: IMM, immunity; RED/STE, redox and oxidoreductive stress; CSR, chemosensory reception; DIG, blood and sugar food digestive; PROT, proteolysis; CYT/STR, cytoskeletal and structural; TRP, transport; R/T/T, replication, transcription, and translation; MET, metabolism; DIV, diverse functions; UNK, unknown functions. The total number of genes in each category is indicated in parentheses.

**Fig. 3.** Orthology and chromosomal synteny among *Ae. aegypti*, *An. gambiae*, and *D. melanogaster*. **(A)** Each circle represents a gene set for *Ae. aegypti* (Ae), *An. gambiae* (An), and *D. melanogaster* (Dm). Because a gene can be involved in several homologies, gene sets do not always have the same number of genes within intersections (e.g., in the *Ae*-*Dm* comparison, 943 *Ae* genes are similar to *Dm* and 925 *Dm* genes are similar to *Ae*). **(B)** *Aedes aegypti* chromosomes are represented in gray (not to scale). Chromosome arms are designated as "p" and "q"; chromosome 1 has no arm distinctions. Colored chromosomes represent the syntenic chromosome from *An. gambiae* or *D. melanogaster* (not to scale). Solid and dashed



conserved evolutionary associations between *Ae. aegypti* and *An. gambiae* or *D. melanogaster* chromosomes (tables S3 and S21). Most of the *Ae. aegypti* chromosome arms, with the exception of 2p and 3q, exhibited a distinct one-to-one correlation with *An. gambiae* and *D. melanogaster* chromosome arms with respect to the proportion of orthologous genes conserved between chromosome arm pairs (Fig. 3B). These findings confirm and extend previous results that compared a small number (~75) of *Ae. aegypti* genes with orthologs in *An. gambiae* and *D. melanogaster* (31).

Maps of conserved local gene arrangements (microsynteny) were computed by identifying blocks of at least two neighboring single-copy orthologs in each pair of genomes and allowing not more than two intervening genes (14). In line with the species divergence times, twice as many orthologs are similarly arranged between these mosquito species than between either of them and the fruit fly (table S22) (32); 1345 microsyntenic blocks were identified between *Ae. aegypti* and *An. gambiae*, containing 5265 out of a total of 6790 single-copy orthologs (tables S5 and S22). When *D. melanogaster* is used as an outgroup to count synteny breaks that have occurred in each mosquito lineage since their radiation, the data indicate a rate of genome shuffling in the *Ae. aegypti* lineage greater by a factor of ~2.5 than that in the *An. gambiae* lineage (14). However, this estimate may be inflated because of the fragmented nature of the current *Ae. aegypti* genome assembly. Thus, the highly repetitive nature of the *Ae. aegypti* genome appears to have facilitated local gene rearrangements, but it does not appear to have had a gross influence on chromosomal synteny.

**Concluding remarks.** The draft genome sequence of *Ae. aegypti* will stimulate efforts to elucidate interactions at the molecular level between mosquitoes and the pathogens they transmit. This already can be seen in, for example, analysis of components of the Toll immune sig-

naling pathway (33) and identification of genes encoding insulin-like hormone peptides (34).

We expect that the sequence data will facilitate the identification of *Ae. aegypti* genes encoding recently described midgut receptors for dengue virus (35). Dengue vector competence is a quantitative trait, and multiple loci determine virus midgut infection and escape barriers (36). Unfortunately, the fragmented nature of the genome sequence and its low gene density have precluded its use in the identification of a comprehensive list of candidate genes for vector competence phenotypes or sex determination. The sequence may be used to improve the resolution of the current genetic map (37) and to integrate transcriptional profiling data with genetic studies (38), but filling gaps in the assembled sequence remains a high priority, especially when exploring genetic variations between the sequenced strain and field populations of *Ae. aegypti*.

The ongoing genome project on *Culex pipiens quinquefasciatus*, a vector for lymphatic filariasis and West Nile virus, will provide additional resources to underpin studies to systematically study common and mosquito species-specific gene function. Such analyses should improve our understanding of mosquito biology and the complex role of mosquitoes in the transmission of pathogens, and may result in the development of new approaches for vector-targeted control of disease.

#### References and Notes

1. B. J. Beaty, W. C. Marquardt, *Biology of Disease Vectors* (Univ. Press of Colorado, Niwot, CO, ed. 1, 1996).
2. S. R. Christophers, *Aedes aegypti* (L.): *The Yellow Fever Mosquito, Its Life History, Bionomics and Structure* (Cambridge Univ. Press, Cambridge, 1960).
3. A. N. Clements, *The Biology of Mosquitoes* (Chapman & Hall, London, 1992).
4. D. W. Severson, S. E. Brown, D. L. Knudson, *Annu. Rev. Entomol.* **46**, 183 (2001).
5. O. Tomori, *Crit. Rev. Clin. Lab. Sci.* **41**, 391 (2004).
6. World Health Organization, *Dengue and Dengue Haemorrhagic Fever* (World Health Organization, Geneva, 2002).

7. B. L. Ligon, *Semin. Pediatr. Infect. Dis.* **17**, 99 (2006).
8. J. S. Mackenzie, D. J. Gubler, L. R. Petersen, *Nat. Med.* **10**, 598 (2004).
9. R. A. Holt et al., *Science* **298**, 129 (2002).
10. E. M. Zdobnov et al., *Science* **298**, 149 (2002).
11. M. W. Gaunt, M. A. Miles, *Mol. Biol. Evol.* **19**, 748 (2002).
12. J. Krzywinski, O. G. Grushko, N. J. Besansky, *Mol. Phylogenet. Evol.* **39**, 417 (2006).
13. G. B. J. Craig, W. A. Hickey, in *Genetics of Insect Vectors of Disease*, J. W. Wright, R. Pal, Eds. (Elsevier, New York, 1967), pp. 67–131.
14. See supporting material on Science Online.
15. D. B. Jaffe et al., *Genome Res.* **13**, 91 (2003).
16. A. M. Warren, J. M. Crampton, *Genet. Res.* **58**, 225 (1991).
17. C. Feschotte, *Mol. Biol. Evol.* **21**, 1769 (2004).
18. M. R. Coy, Z. Tu, *Insect Mol. Biol.* **14**, 537 (2005).
19. N. Craig, R. Cragie, M. Gellert, A. Lambowitz, Eds., *Mobile DNA II* (American Society for Microbiology Press, Washington, DC, 2002).
20. J. M. Tubio, H. Naveira, J. Costas, *Mol. Biol. Evol.* **22**, 29 (2005).
21. I. R. Arkhipova, K. I. Pyatkov, M. Meselson, M. B. Evgen'ev, *Nat. Genet.* **33**, 123 (2003).
22. X. Zhang, N. Jiang, C. Feschotte, S. R. Wessler, *Genetics* **166**, 971 (2004).
23. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
24. L. C. Bartholomay et al., *Infect. Immun.* **72**, 4114 (2004).
25. H. R. Sanders et al., *Insect Biochem. Mol. Biol.* **35**, 1293 (2005).
26. H. Ranson, personal communication.
27. A. Wise, K. Gearing, S. Rees, *Drug Discov. Today* **7**, 235 (2002).
28. Q. Ren, K. H. Kang, I. T. Paulsen, *Nucleic Acids Res.* **32**, D284 (2004).
29. C. Schutt, R. Nothiger, *Development* **127**, 667 (2000).
30. O. Marinotti et al., *Insect Mol. Biol.* **15**, 1 (2006).
31. D. W. Severson et al., *J. Hered.* **95**, 103 (2004).
32. E. M. Zdobnov, P. Bork, *Trends Genet.* **23**, 16 (2007).
33. S. W. Shin, G. Bian, A. S. Raikhel, *J. Biol. Chem.* **281**, 39388 (2006).
34. M. A. Riehle, Y. Fan, C. Cao, M. R. Brown, *Peptides* **27**, 2547 (2006).
35. R. F. Mercado-Curiel et al., *BMC Microbiol.* **6**, 85 (2006).
36. C. F. Bosio, R. E. Fulton, M. L. Salasak, B. J. Beaty, W. C. Black, *Genetics* **156**, 687 (2000).
37. D. W. Severson, J. K. Meece, D. D. Lovin, G. Saha, I. Morlais, *Insect Mol. Biol.* **11**, 371 (2002).
38. R. C. Jansen, J. P. Nap, *Trends Genet.* **17**, 388 (2001).
39. The *Aedes aegypti* genome sequencing project at the microbial sequencing centers and VectorBase was funded by National Institute of Allergy and Infectious Diseases (NIAID) contracts HHSN266200309D266030071,

HHSN266200400001C, and HHSN266200400039C and was supported in part by NIAID grants U01 AI50936 (D.W.S.), R01 AI059492 (A.S.R., G.D.), 5 R01 AI61576-2 (G.D.), and R37 AI024716 (A.S.R.) and by Swiss National Science Foundation grant SNF 3100AO-112588/1 (E.M.Z.). We acknowledge the excellent work of the Broad Genome Sequencing Platform and the Venter Institute Joint Technology Center. We thank C. Town, N. Hall, and E. Kirkness for critical comments and the *Aedes aegypti*

research community for their enthusiastic support and willing assistance in this project. On 1 October 2006 The Institute for Genomic Research merged with the J. Craig Venter Institute. The *Ae. aegypti* genome can also be accessed at VectorBase (<http://aegypti.vectorbase.org>).

#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/1138878/DC1](http://www.sciencemag.org/cgi/content/full/1138878/DC1)  
Materials and Methods

Figs. S1 to S7  
Tables S1 to S23  
References

15 December 2006; accepted 7 May 2007  
Published online 17 May 2007;  
10.1126/science.1138878  
Include this information when citing this paper.

## REPORTS

# Do Vibrational Excitations of CHD<sub>3</sub> Preferentially Promote Reactivity Toward the Chlorine Atom?

Shannon Yan,<sup>1</sup> Yen-Tien Wu,<sup>1</sup> Bailin Zhang,<sup>1\*</sup> Xian-Fang Yue,<sup>1†</sup> Kopin Liu<sup>1,2‡</sup>

The influence of vibrational excitation on chemical reaction dynamics is well understood in triatomic reactions, but the multiple modes in larger systems complicate efforts toward the validation of a predictive framework. Although recent experiments support selective vibrational enhancements of reactivities, such studies generally do not properly account for the differing amounts of total energy deposited by the excitation of different modes. By precise tuning of translational energies, we measured the relative efficiencies of vibration and translation in promoting the gas-phase reaction of CHD<sub>3</sub> with the Cl atom to form HCl and CD<sub>3</sub>. Unexpectedly, we observed that C–H stretch excitation is no more effective than an equivalent amount of translational energy in raising the overall reaction efficiency; CD<sub>3</sub> bend excitation is only slightly more effective. However, vibrational excitation does have a strong impact on product state and angular distributions, with C–H stretch-excited reactants leading to predominantly forward-scattered, vibrationally excited HCl.

Several decades of experimental and theoretical molecular collision studies culminated in the formulation of Polanyi's rules of reaction dynamics (1). For reactions of an atom with a diatomic molecule, the rules predict the efficiency of reactant vibrational and translational energy in driving reactions over barriers; namely, vibration can be more effective than translation for a barrier located late along the reaction coordinate, and the reverse is true for reactions with early barriers. An extension of the rules to reactions of polyatomic species becomes ambiguous as a result of the higher degrees of freedom associated with multiple types of vibrational motion. Thus, one may ask: Are different vibrational modes equivalent in their capacity to promote a polyatomic reaction?

In recent years, the issue of mode-specific or bond-selective chemistry (2–5) has been the sub-

ject of several pioneering investigations, for which the reaction of the Cl atom with methane is becoming the benchmark (6–19). For example, Simpson *et al.* found that one-quantum excitation in the antisymmetric stretch ( $\nu_3$ ) mode of CH<sub>4</sub> increases the reaction rate by a factor of ~30 (10). On the other hand, Zhou *et al.* observed a mere threefold reactivity enhancement for one-quantum excitation of bending ( $\nu_4$ ) or torsional ( $\nu_2$ ) modes of CH<sub>4</sub> and CD<sub>4</sub> (18), in contrast to 200-fold and 80-fold enhancements measured earlier (12, 13). Further experiments (17) and a quasiclassical trajectory calculation (20) supported the results of Zhou *et al.* Moreover, Yoon *et al.* found that excitation of the  $\nu_1 + \nu_4$  symmetric stretch-bend combination mode of CH<sub>4</sub> enhances reactivity toward the Cl atom roughly twice as much as does the nearly isoenergetic excitation of the antisymmetric combination  $\nu_3 + \nu_4$ , which itself promotes a 10-fold rate enhancement over ground-state methane (6). In a similar study, Yoon *et al.* observed a sevenfold reactivity increase of CH<sub>3</sub>D when the symmetric, rather than antisymmetric, C–H stretching mode was initially excited (8). All these experiments, however, were performed at a fixed translational or collision energy ( $E_c$ ); thus, the enhanced reactivity refers to a comparison with the ground-state reaction at the same  $E_c$ . As elegant as these experiments are, it remains

uncertain whether vibrational motion is more effective in driving this reaction than translation.

We report here a series of experiments aimed to resolve this uncertainty for the Cl + CHD<sub>3</sub> → HCl + CD<sub>3</sub> reaction. We first studied the ground-state reaction over a wide energy range from the threshold to about 20 kcal/mol of excess energy. Experiments were then performed for the reaction with C–H stretch-excited CHD<sub>3</sub>, again over a range of initial  $E_c$ . To refine the comparison, we also present the results for the bend- and/or torsion-excited reactants. We performed all measurements under single-collision conditions, using the rotatable, crossed molecular-beam apparatus described previously (21, 22). The Cl beam was generated by a pulsed high-voltage discharge of ~4% Cl<sub>2</sub> seeded in a pulsed supersonic expansion of either Ne or He at 6 atm. The CHD<sub>3</sub> beam was also produced by pulsed supersonic expansion of either pure CHD<sub>3</sub> or ~20% CHD<sub>3</sub> seeded in H<sub>2</sub> (for acceleration) at 5 atm. Both beams were collimated by double skimmers and crossed in a differential-pumped scattering chamber.  $E_c$  was tuned by varying the intersection angle of the two molecular beams. A pulsed ultraviolet laser that was operated near 333 nm probed the ground-state CD<sub>3</sub> product via (2 + 1) resonance-enhanced multiphoton ionization, and a time-sliced velocity imaging technique mapped the recoil vector of the CD<sub>3</sub><sup>+</sup> ion (21). For studies with C–H stretch-excited reactants, an infrared (IR) laser was used to excite CHD<sub>3</sub> directly in front of the first skimmer (19). For reactions with bend-excited reactants, a heated pulsed valve for thermal excitation was used instead (18).

Figure 1 shows two typical raw images, with and without the IR-pumping laser, of the probed CD<sub>3</sub>( $v = 0$ ) products at  $E_c = 8.9$  kcal/mol. Superimposed on the images are the scattering directions; the 0° angle refers to the initial CHD<sub>3</sub> beam direction in the center-of-mass frame. Thanks to the time-sliced velocity imaging approach, even the raw data can be easily interpreted by inspection. Whereas the IR-off image is dominated by a side-scattered structure, the IR-on image exhibits two distinct ringlike features reflecting the impact of C–H stretch excitation on the reaction dynamics (23). A sharp forward peak now appears in the inner ring, and additional broad-scattered products form the outer ring. The

<sup>1</sup>Institute of Atomic and Molecular Sciences, Academia Sinica, Post Office Box 23-166, Taipei, Taiwan 10617.

<sup>2</sup>Department of Chemistry, National Taiwan University, Taipei, Taiwan 10617.

\*Present address: Department of Chemistry, Wayne State University, Detroit, MI 48202, USA.

†Present address: Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China.

‡To whom correspondence should be addressed. E-mail: [kliu@po.iam.s.sinica.edu.tw](mailto:kliu@po.iam.s.sinica.edu.tw)