Anopheles gambiae pilot gene discovery project: Identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines

George Dimopoulos*, Thomas L. Casavant†, Shereen Chang†, Todd Scheetz†, Chad Roberts†, Micca Donohue†, Jörg Schultz*, Vladimir Benes*, Peer Bork*, Wilhelm Ansorge*, Marcelo Bento Soares†, and Fotis C. Kafatos*‡

*European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany; and †University of Iowa, 451 Eckstein Medical Research Building, Iowa City, IA 52242

Contributed by Fotis C. Kafatos, March 31, 2000

Together with AIDS and tuberculosis, malaria is at the top of the list of devastating infectious diseases. However, molecular genetic studies of its major vector, *Anopheles gambiae*, are still quite limited. We have conducted a pilot gene discovery project to accelerate progress in the molecular analysis of vector biology, with emphasis on the mosquito's antimalarial immune defense. A total of 5,925 expressed sequence tags were determined from normalized cDNA libraries derived from immune-responsive hemocyte-like cell lines. The 3,242 expressed sequence tag-containing cDNA clones were grouped into 2,380 clone clusters, potentially representing unique genes. Of these, 1,118 showed similarities to known genes from other organisms, but only 27 were identical to previously known mosquito genes. We identified 38 candidate genes, based on sequence similarity, that may be implicated in immune reactions including antimalarial defense; 19 of these were shown experimentally to be inducible by bacterial challenge, lending support to their proposed involvement in mosquito immunity.

Transmission of the malaria parasite, *Plasmodium*, involves two complex and obligatory life cycles in the vector mosquito as well as in the human host. Interruption of either cycle would attenuate the spread of the disease. The prospect of control strategies based on transmission blocking in the vector (1, 2) has energized studies on the molecular genetics of *Anopheles gambiae*. Special attention has been directed toward the main organs with which the parasite interacts during its development in the mosquito, the midgut and the salivary glands (3–6). The observation that the parasite is destroyed completely in refractory mosquito strains but also sustains substantial losses in fully susceptible mosquitoes (7) has recently drawn attention to the study of the mosquito's innate immune system (8–15). Immune reactions induced by malaria infection correlate with the life cycle of *Plasmodium* in the vector mosquito; they have been demonstrated at the molecular level in the midgut and salivary gland epithelia, in hemocytes, and in the fat body, a liver analogue in insects (11).

Difficulties in rearing malaria mosquitoes under laboratory conditions and the limited amount of biological material that can be obtained from mosquito organs are obstacles to the isolation of *A. gambiae* genes. Gene cloning from this organism started nearly a decade ago and has generated only ≈450 putative coding sequences in the public protein databases. Massive sequencing of cDNAs from source-specific libraries of other organisms has proven to be a powerful approach to gene discovery (16). Putative functions can be proposed for the discovered genes either through homology searches of global databases or by mass expression profiling with the recently developed cDNA microarray technologies (17). A powerful stimulus to the study of parasites has already resulted from genomics projects, including the expressed sequence tag (EST) projects of the worms *Brugia malayi* and *Schistosoma sp.* and the genome sequencing projects of the protozoa *Plasmodium falciparum* and *Leishmania major* (18–21). In a pilot attempt to evaluate the efficiency of mass cDNA sequencing for gene discovery in *A. gambiae*, ESTs were generated from random clones of normalized cDNA libraries. Our special interest in immunity genes led us to the choice of recently established immune responsive hemocyte-like cell lines, which are known to express high levels of various immune markers including antimicrobial peptides, putative recognition molecules, serine proteases and their inhibitors (serpins), and prophenoloxidases (PPO) (refs. 10, 15, and 22 and A. Danielli, personal communication). A significant number of clone clusters were similar to genes that encode proteins known to operate in invertebrate and vertebrate innate defense mechanisms; half of them were shown experimentally to be immune responsive.

Methods

Cell Cultures and Immune Challenge. The previously described cell lines 4A-3A and 4A-3B were cultured in Schneider (Sigma) medium supplemented with 10% (vol/vol) BSA at 27°C as described (15) and harvested at a confluent growth phase. The 4A-3B cell line was challenged with 10 μg/ml lipopolysaccharide (Sigma) 6 h before harvest.

RNA Extraction and Library Construction. Extraction of mRNA was performed with the Oligotex Direct mRNA Maxi kit (Qiagen, Chatsworth, CA), and 7 μg of mRNA was used for construction of the cDNA libraries. cDNAs were cloned directionally into a phagemid vector (pT7T3-Pac), and the libraries were normalized as described (23).

Sequencing. The plasmid libraries were electroporated into a DH10-α *Escherichia coli* strain, and DNA extracted from randomly selected clones was subjected to automated sequencing.

Abbreviations: PPO, prophenoloxidase; EST, expressed sequence tag.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AJ280015–AJ285940).

‡To whom reprint requests should be addressed. E-mail: kafatos@embl-heidelberg.de.

Table 1. EST statistics

Analyzed data	4A-3A	4A-3B	Total
ESTs	3,269	2,656	5,925
Clones	1,839	1,403	3,242
Average insert size, bp			1,500
Average EST length, bp			375
EST clusters			4,122
Clone clusters			2,380
Homologous clone clusters			1,118
Identical to *Anopheles*			27
Potential immunity genes			38

Numbers of sequenced cDNA clones and generated ESTs from the libraries constructed from cell lines 4A-3A and 4A-3B. The average insert size was calculated for 100 cDNA clones from each library, and EST length was calculated from the total set of 5,925 ESTs. ESTs with 97% or greater identity over a 100-bp region were clustered together forming 4,122 EST clusters. Clusters including the 5′ and 3′ end sequences of the same clone were grouped together forming 2,380 clone clusters, each potentially representing an individual gene. One or more ESTs of 1,118 clone clusters had a significant BLASTX *E* value ($<10^{-4}$) to other proteins in a nonredundant SWISSPROT and SPTREMBL database (24); a small number of these seem to be chimeric. A total of 27 clone clusters had protein sequences identical to those of *A. gambiae* genes in the database, and 38 clone clusters were similar to genes known to play potential roles in innate immunity.

Sequence Analysis. ESTs were checked for vector sequence contaminants. Sequence analysis against databases was performed with the BLASTX software (24). The clone sequences were subjected to an all-against-all sequence comparison where clones sharing at least one EST with 97% or greater identity over a 100-bp region were grouped together in the same cluster. The database entry keywords of the homologue sequences were used for the grouping of clone clusters in functional classes with a modified version of the EUCLID software (25, 26).

Reverse Transcription–PCR Expression Assays. Expression levels of selected ESTs were assayed in naïve and bacteria-challenged 4A-3A and 4A-3B cell lines by reverse transcription–PCR as described (15). The following primers were used for the expression analysis: I.2a, 5′-TGGATGGTATCGGGTTCCG-3′; I.2b, 5′-GGGATGGACGACAATCTCC-3′; I.4a, 5′-TCATAGCGGAACGATGGGC-3′; I.4b, 5′-GGAGGTGTAGATGCCCGGA-3′; I.6a, 5′-CTGCTCACTTGTATCGGGC-3′; I.6b, 5′-TATCTACTACCCGCTGCGC-3′; I.7a, 5′-TCCGGTGGACCGCTGATG-3′; I.7b, 5′-CATTGTAGAAAGCATATC-3′; I.11a, 5′-CCCAGCCGCAACTGCAGCC-3′; I.11b, 5′-GCTGGATCGTAGATCGTGC-3′; II.4a, 5′-CGGTAGACAAATCGATGGC-3′; II.4b, 5′-CGGTTAGCGAGTGGTCGGG-3′; II.5a, 5′-GGGTCGAGCTTCGCACCG-3′; II.5b, 5′-AAATCCAATCTCCCTTC-3′; II.6a, 5′-GGGGTTCACAGGAATTACT-3′; II.6b, 5′-TCATCAAGGACACTTGGGG-3′; II.7a, 5′-CAACGGTGACTTCTACTGG-3′; II.7b, 5′-GTGCCAGCACGATATTACC-3′; II.8a, 5′-GCCCAAGTACGACCACACC-3′; II.8b, 5′-GGACAGCAGGTCTCACTCG-3′; II.10a, 5′-CGGTGTGCCACTGTTCGGG-3′; II.10b, 5′-CCGTTTGCCACACTTGCCC-3′; II.11a, 5′-CGTCAGCTAGCCGCACTGC-3′; II.11b, 5′-TCGTGCTGTGGTAATCCGG-3′; II.12a, 5′-GGAGTACGAGTCGGGCGGG-3′; II.12b, 5′-CGAGTAATGGTACCCACGG-3′; II.14a, 5′-GTGCAAGAAGACCCCATCG-3′; II.14b, 5′-TTCCATACCACCATGCCCC-3′; II.15a, 5′-TTAAATCTGTATGTCTGCC-3′; and II.15b, 5′-TAACACGATGCTCAGCTGC-3′.

Results and Discussion

Source Libraries and EST Sequencing. Two cell lines with overlapping but distinct immune expression profiles, 4A-3A and 4A-3B, were used as the starting material. The 4A-3B cell line expresses high levels of PPO transcripts, and 4A-3A expresses strongly various other immune markers (15). The latter line, 4A-3B, was challenged with lipopolysaccharide, a potent bacterial immune elicitor, for 8 h before mRNA extraction for enrichment of immune gene transcripts (10, 15). Normalized, directionally cloned poly(T)-primed cDNA libraries were constructed as described (23), and randomly selected clones were sequenced. The average insert size was estimated as 1.5 kilobases by PCR amplification of inserts from 100 randomly selected clones.

A total of 3,242 clones were sequenced, mostly from both ends, generating 5,925 ESTs with an average length of 375 bp (range: 22–1,114 bp; 375 × 5,925 = ≈2.2 megabases of total sequence generated; Table 1). The generated ESTs corresponded to 2,380 clone clusters potentially representing individual genes, suggesting that the overall redundancy of the libraries may be only ≈27%. However, failure to detect overlaps between partial cDNA clones cannot be excluded. Normalization of the libraries did suppress the highly abundant messages and consequently increased the number of discovered genes.

Sequence Similarities. BLASTX analysis comparing all clone clusters against a nonredundant database generated from SWISSPROT and SPTREMBL (25) revealed that 1,118 clone clusters (47% of the total) are significantly similar (E value $< 10^{-4}$) to known genes. Of these, 57 clusters showed the highest similarity to other known *A. gambiae* genes, but only 27 were identical. Thus, the vast majority of the clone clusters can be considered putative, previously unidentified *A. gambiae* genes. Of the clone clusters that showed significant BLASTX hits, only 99 (8.9%) showed similarity to known insect genes alone. For 45.7% of the clone clusters, similarities were comparable for both insect and mammalian sequences. For 36.4%, they were highest to mammalian

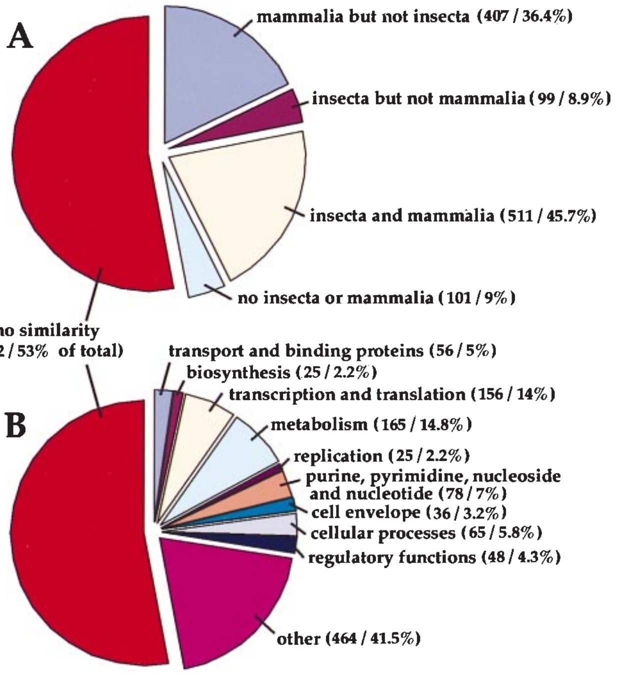

Fig. 1. Distribution of clone clusters in gene classes based on BLASTX E values (24). Of the 2,380 clone clusters, 1,262 (53%) did not show significant similarity ($E < 10^{-4}$) to genes in the nonredundant SWISSPROT and SPTREMBL databases. The remaining 1,118 (47%) clone clusters with significant BLASTX E values are distributed in the gene classes based on their detected similarities. (*A*) Numbers and percentages of clone clusters with lowest BLASTX E values to genes from mammalia but not insecta, insecta but not mammalia, insecta and mammalia, and neither insecta nor mammalia. (*B*) Numbers and percentages of clone clusters with significant BLASTX E values to known genes belonging to the functional classes as defined earlier (25).

Table 2. Putative serine proteases and serpins

Cluster code	Nucleotide	EST	E value	Homology	Putative function
I.1	319	4A3B-aad-d-12-F	1e-19	Serine protease 14D (*A.g.*)	Activation of PPO cascade
	206	4A3B-aad-d-12-R	1e-05	Easter (*D.m.*)	
	245	4A3B-aax-c-11-F	2e-06	Serine protease 14D (*A.g.*)	
	304	4A3B-aax-c-11-R	3e-06	Serine protease 14D (*A.g.*)	
	366	4A3A-aaq-c-04-F	3e-20	Serine protease 14D (*A.g.*)	
	356	4A3A-aaq-c-04-R	8e-12	Serine protease 14D (*A.g.*)	
	288	4A3A-aat-a-03-F	5e-19	Serine protease 14D (*A.g.*)	
		4A3A-aat-a-03-R	3e-12	Serine protease 14D (*A.g.*)	
I.2	342	4A3A-aal-a-11-F	5e-17	Coagulation factor D (*T.t.*)	Coagulation cascade
	338	4A3A-aal-a-11-R	6e-05	Trypsin like serine protease (*C.f.*)	
	379	4A3A-P2G4-R	5e-12	Masquerade like protein (*P.l.*)	
I.3	340	4A3A-aam-e-02-F	2e-05	Easter (*D.m.*)	Developmental
	241	4A3A-aam-e-02-R	4e-07	Factor C (*C.r.*)	Coagulation cascade
I.4	181	4A3B-aaa-a-02-F	3.0	Trypsin-like (*C.f.*)	Complement
	277	4A3B-aaa-a-02-R	4e-10	Trypsin-like (*C.f.*)	
I.5	301	4A3A-aal-c-02-R	2e-05	Hemocyte protease-1 (*M.s.*)	Serine protease
I.6	391	4A3A-aaq-f-04-R	0.003	Serine protease 14D (*A.g.*)	Serine protease
	447	4A3A-aaq-f-04-R	6e-15	Elastase precursor (*P.o.*)	
I.7	238	4A3B-aad-b-06-F	4e-15	Serine protease 14D (*A.g.*)	Activation of PPO cascade
I.8	285	4A3B-aau-g-07-R	Identical	Serine protease 22D (*A.g.*)	Adhesive serine protease
I.9	431	4A3B-aau-c-06-F	1e-04	Serpin (*P.l.*)	Coagulation factor inhibitor
	274	4A3B-aau-c-06-R	8e-10	Serpin (*P.l.*)	
	356	4A3B-aag-b-09-R	4e-05	Serpin (*P.l.*)	
	637	4A3A-P4E1-F	2e-09	Serpin (*P.l.*)	
	411	4A3A-P4E1-R	0.002	Serpin (*R.n.*)	
I.10	513	4A3A-aap-b-5-F	Identical	Ag.serpin (*A.g.*)	Serine protease inhibitor
	412	4A3A-aap-b-5-R	Identical	Ag.serpin (*A.g.*)	
	557	4A3A-aas-g-03-F	Identical	Ag.serpin (*A.g.*)	
	416	4A3A-aas-g-03-R	Identical	Ag.serpin (*A.g.*)	
I.11	594	4A3A-P6H1-F	1e-12	Serpin (*H.c.*)	Serine protease inhibitor
	448	4A3A-P6H1-R	1e-05	Serpin (*A.t.*)	
I.12	430	4A3A-aap-b-10-F	4e-08	Antiplasmin serpin (*B.t.*)	Serine protease inhibitor

Clusters with similarity to putative immunity-related genes (note that the following legend applies to Tables 3 and 4 as well). The clusters were assigned short numerical codes and are displayed in the same order as presented in the main text. Their EST identifiers indicate the cell line origin (4A-3A or 4A-3B), the position in the plates of arrayed cDNA libraries, and finally the vector primer used for sequencing [reverse (R) and forward (F)]. In the great majority of clones, the R primer is near the 5′ end of the ESTs, and the F primer is near the 3′ end. The BLASTX hits with the lowest *E* values (implying the most significant similarities) are indicated in the table; inference of putative function is derived from these and additional protein matches with highly significant similarities. Previously known *A. gambiae* genes are marked "identical." The organisms corresponding to listed similar proteins are indicated with italics initials in the homology column: *A.g., A. gambiae; D.m., Drosophila melanogaster; T.t., Tachypleus tridentiatus; C.f., Ctenocephalides felis; C.r., Carcinoscorpius rotundicauda; M.s., Manduca sexta; P.o., Paralichthys olivaceus; P.l., Pacifastacus leniusculus; R.n., Rattus norvegicus; H.c., Hyalophora cecropia; A.t., Arabidopsis thaliana; B.t., Bos taurus.*

sequences, and for 9%, they were highest to genes from other organisms (Fig. 1*A*). The similarities suggested putative functions for 654 clone clusters, which were grouped in nine distinct functional groups (26) as indicated in Fig. 1*B*.

Potential Immunity Genes. As many as 38 clone clusters showed significant similarities to classes of known innate immunity genes, reflecting the origin of the cDNAs from hemocyte-like cell lines (15). Blood cells from both vertebrates and invertebrates are known to play key roles in defensive innate immunity mechanisms, such as melanization leading to encapsulation, coagulation and complement cascades, phagocytosis, and production of antimicrobial peptides. These 38 clusters are discussed below in the order of their presentation in Tables 2–4. For each cluster, inference of putative function was based on consideration of several matches with high scores. The encoded gene products were classified into three broad groups, as follows.

Putative Serine Proteases and Serpins. Seven clusters (I.1–I.7) encode putative homologues of serine proteases that bear "clip-domains," a common feature of regulatory immunity proteases (27). Notably, similarities were detected to the *A. gambiae* immune responsive 14D serine protease (28), to clotting, to coagulation, to complement factors, and to PPO-activating enzymes. One cluster (I.8) is identical to the chitin-binding domain of a multidomain, modular immune responsive serine protease of *A. gambiae* characterized by others (29, 30). Four clusters (I.9–I.12) encode putative serpins with similarities to those of mammals, insects, and other invertebrates, including a coagulation inhibitor. One cluster (I.10) is identical to a previously isolated mosquito serpin (GenBank accession nos. AJ271352 and AJ271353). Serine proteases and their inhibitors are of interest as potential components of the regulated cascade/amplification reactions of blood clotting, complement, and other immune responses. Examples are the known regulators of PPO and coagulation cascades that have been isolated and characterized from the moths *Bombyx mori* and *Manduca sexta* (27, 31).

Putative Adhesive Proteins. Proteins encoded by 15 clone clusters resemble adhesive proteins, including proteins capable of recognizing and binding to microorganisms. Two of these (II.1 and II.2) resemble lectins, including a rat intracellular mannose binding lectin (32) and an immune-inducible galactose binding lectin of the mosquito (6, 10). Lectins are involved frequently in

Table 3. Putative adhesive proteins

Cluster code	Nucleotide	EST	E value	Homology	Putative function
II.1	451	4A3B-aae-c-10-R	2e-05	P58 lectin (*R.n.*)	Adhesion
II.2	397	4A3A-aao-g-02-R	8e-11	Galactose lectin (*A.g.*)	Adhesion
II.3	403	4A3B-aau-c-07-F	Identical	GNBP (*A.g.*)	Binding to bacteria
	368	4A3B-aau-c-07-R	Identical	GNBP (*A.g.*)	
II.4	743	4A3A-P11B10-R	2e-19	β-1,3 glucanase (*S.p.*)	Binding to bacteria
	743	4A3A-P11B10-R	6e-19	GNBP (*H.c.*)	
II.5	400	4A3A-P4E10-F	4e-10	GNBP (*B.m.*)	Binding to bacteria
II.6	332	4A3B-aaa-d-04-F	2e-04	PGRP (*T.n.*)	PPO cascade inhibitor
II.7	324	4A3A-P2B6-F	3e-04	Hemomucin (*D.m.*)	Opsoninization
	210	4A3A-P2B6-R	0.009	Hemomucin (*D.m.*)	
	326	4A3B-aau-b-09-F	1e-05	Hemomucin (*D.m.*)	
	355	4A3B-aau-b-09-R	7e-27	Hemomucin (*D.m.*)	
	422	4A3A-aao-f-07-F	3e-10	Hemomucin (*D.m.*)	
	386	4A3A-aao-f-07-R	9e-33	Hemomucin (*D.m.*)	
II.8	431	4A3A-P8G3-F	2e-10	dSR-C1 (*D.m.*)	Adhesion, recognition
	181	4A3A-P8G3-R	0.053	dSR-C1 (*D.m.*)	
II.9	565	4A3A-P2E1-R	9e-34	CD36 (*A.g.*)	Apoptotic cell phagocytosis
II.10	402	4A3B-aaw-d-03-R	4e-23	Endochitinase precursor (*M.s.*)	Chitin binding
II.11	448	4A3B-aaf-e-09-F	3e-13	Ficolin (*S.s.*)	Adhesion, phagocytosis
		4A3B-aaf-e-09-R	9e-14	Angiopoietin Y1 (*H.s.*)	
II.12	358	4A3B-aag-c-08-F	1e-16	Tachylectin (*T.t.*)	Adhesion, phagocytosis
	391	4A3B-aag-c-08-R	0.004	Angiopoietin (*B.t.*)	
II.13	551	4A3A-P4A6-F	3e-19	Angiopoietin Y1 (*H.s.*)	Adhesion, phagocytosis
II.14	517	4A3A-aak-e-08-R	4e-45	Ficolin (*S.s.*)	Adhesion, phagocytosis
II.15	320	4A3B-aaj-e-03-R	2e-25	Angiopoietin related (*M.m.*)	Adhesion, phagocytosis
	283	4A3B-aav-c-11-F	3e-15	Ficolin (*S.s.*)	

Abbreviations have been defined in the legend to Table 2, with the additions of GNBP, Gram-negative bacteria binding protein; PGRP, peptidoglycan recognition protein; *S.p.*, *Strongylocentrotus purpuratus*; *B.m.*, *Bombyx mori*; *T.n.*, *Trichoplusia ni*; *S.s.*, *Sus scrofa*; *H.s.*, *Homo sapiens*; *M.m.*, *Mus musculus.*

Table 4. Other putative immunity proteins

Cluster code	Nucleotide	EST	E value	Homology	Putative function
III.1	630	4A3A-P3A8-R	2e-04	Cecropin A (*Ae.al.*)	Antimicrobial
III.2	443	4A3B-aae-f-01-R	Identical	AgIRSP (*A.g.*)	Unknown
	395	4A3A-aao-b-11-F	Identical	AgIRSP (*A.g.*)	
		445	4A3A-aao-b-11-R	Identical	AgIRSP (*A.g.*)
III.3	248	4A3A-aar-h-04-F	4e-06	α-2-macroglobulin (*L.j.*)	Complement
III.4	473	4A3A-aao-h-01-F	2e-65	Pelle-associated protein (*D.m.*)	Toll pathway/signalling
	309	4A3A-aao-h-01-R	1e-48	Pelle-associated protein (*D.m.*)	
III.5	438	4A3B-aah-f-01-R	1e-09	Cactus (*D.m.*)	Toll pathway/signalling
	582	4A3A-aat-a-11-F	5e-09	Cactus (*D.m.*)	
	463	4A3A-aat-a-11-R	4e-40	Cactus (*D.m.*)	
III.6		4A3A-aak-h-10-R	9e-13	κ-B binding protein (*M.m.*)	κ-B element binding
III.7	496	4A3B-aaj-e-04-F	Identical	AgPPO5 (*A.g.*)	Melanization
	436	4A3B-aaj-e-04-R	Identical	AgPPO5 (*A.g.*)	
	402	4A3A-aak-d-12-F	Identical	AgPPO5 (*A.g.*)	
	516	4A3A-aak-d-12-R	Identical	AgPPO5 (*A.g.*)	
III.8	400	4A3A-aap-h-01-F	Identical	AgPPO2 (*A.g.*)	Melanization
	359	4A3A-aap-h-01-R	Identical	AgPPO2 (*A.g.*)	
III.9	365	4A3B-aax-g-09-R	Identical	AgIRP (*A.g.*)	Iron metabolism
	487	4A3A-aay-g-05-R	Identical	AgIRP (*A.g.*)	
	379	4A3A-aak-f-02-F	Identical	AgIRP (*A.g.*)	
	225	4A3A-abc-d-03-R	Identical	AgIRP (*A.g.*)	
	380	4A3A-P2C6-F	Identical	AgIRP (*A.g.*)	
	513	4A3B-P1D11-R	Identical	AgIRP (*A.g.*)	
	434	4A3B-aah-d-06-R	Identical	AgIRP (*A.g.*)	
III.10	403	4A3B-aaj-a-03-F	2e-06	Ferritin G (*C.e.*)	Iron metabolism
	356	4A3B-aaj-a-03-R	4e-10	Ferritin G (*C.e.*)	
	456	4A3A-aas-h-10-F	8e-05	Ferritin G (*C.e.*)	
	458	4A3A-aas-h-10-R	1e-12	Ferritin G (*C.e.*)	
III.11	338	4A3B-aaa-d-03-F	6e-21	Ferritin HCH (*Ae.a*)	

Abbreviations have been defined in the legends to Tables 2 and 3 with the additions of *Ae.al.*, *Aedes albopictus*; *L.j.*, *Lamperta japonica*; *C.e.*, *Calpodes ethlius*; *Ae.a.*, *Aedes aegypti.*

opsonization and aggregation of microorganisms through their carbohydrate-binding domains (33). Three distinct mosquito homologues of the *B. mori* GNBP, one of them previously isolated and characterized in *A. gambiae* as an immune marker (10), are represented by the clusters II.3–II.5. GNBPs show similarities to the β-1,3 glucan-binding region of glucanases and are likely components of the PPO-activation cascade (31), as is PGRP, which is highly similar to cluster II.6. The PPO cascades can be triggered by diverse microbial surface components such as lipopolysaccharide, β-1,3 glucan, and peptidoglycan; the latter moiety is believed to trigger the PPO cascade in *B. mori* through binding to PGRP (34). The products of clusters II.7 and II.8 resemble domains found, respectively, in hemomucin (35), the putative *Drosophila* opsonin, and the multidomain *Drosophila* scavenger receptor C1 (36). Cluster II.9 encodes the previously isolated mosquito CD36 (accession no. Q17012), a homologue of the *Drosophila* croquemort protein that is involved in phagocytosis of apoptotic cells (37). Cluster II.10 is similar to chitin-binding domains of diverse proteins such as chitinases, mucins, and peritrophins. Finally, clusters II.11–II.15 encode proteins with putative microorganism-binding fibrinogen-like domains. Such domains have been encountered previously in two additional infection-responsive *A. gambiae* genes (G.D., unpublished material), a crab innate immunity lectin that can agglutinate bacteria and enhance defensin activity, and the vertebrate putative phagocytosis mediators, the ficolins (38, 39).

Other Putative Immune Proteins. Cluster III.1 encodes a putative mosquito antimicrobial peptide, cecropin (40), different from that characterized by others (22). One cluster, III.2, corresponds to a recently isolated *A. gambiae* infection-responsive peptide gene of unknown function (accession no. AJ237664). III.3 encodes a member of the complement/α-2-macroglobulin family, other members of which are immune-responsive in *Anopheles* and *Drosophila* (M. Lageux, E. Levashina, and L. Moita, personal communication). Clusters III.4–III.6 encode proteins potentially involved in intracellular immune signaling pathways (41) including putative homologues of the Pelle-associated protein Pellino, of the IκB-like Cactus factor, and of an NFκB motif-binding phosphoprotein (42). Clusters III.7 and III.8 correspond to the previously characterized PPO2 and PPO5 genes (15, 43). Finally, clusters III.9–III.11 correspond to components involved in iron metabolism and regulation: IRP (iron regulatory protein) and ferritin, both of which are implicated in immunity (44–46).

Infection Responsiveness. The 38 putative immune-related clone clusters were subjected to an experimental test of their response to immune challenge. Cell lines 4A-3A and 4A-3B were cocultured for 8 h with heat-killed bacteria, and a reverse transcription–PCR assay was used to detect changes in mRNA prevalence. Indeed, transcripts of one previously known (II.3; not shown) and 18 not previously examined clusters were immune induced, and one, with similarity to domains found in the putative *Drosophila* scavenger receptor homologue, was repressed by exposure to heat-killed bacteria (Fig. 2). It is notable that some members belonging to the same protein family are inducible, and others are not. For example, of the eight putative serine proteases, only five are inducible. Similarly, one of four putative serpins and four of five fibrinogen-like domain proteins are inducible. Differential induction specificities between the two cell lines were noted for some genes, e.g., cluster I.8 is up-regulated by bacterial challenge in cell line 4A-3A but not in 4A-3B. The present number of immune-inducible clone clusters is likely to be an underestimate of the prevalence of immune-related sequences in the collection. Some proteins that are known to be implicated in defense reactions are translationally rather than transcriptionally regulated (e.g., PPO and IRP); they may also be synthesized constitutively and released from the cell or posttranslationally activated on microbial challenge (27). Furthermore, the 1,262 clone clusters that as yet showed no similarity to database entries have not been examined for inducibility.

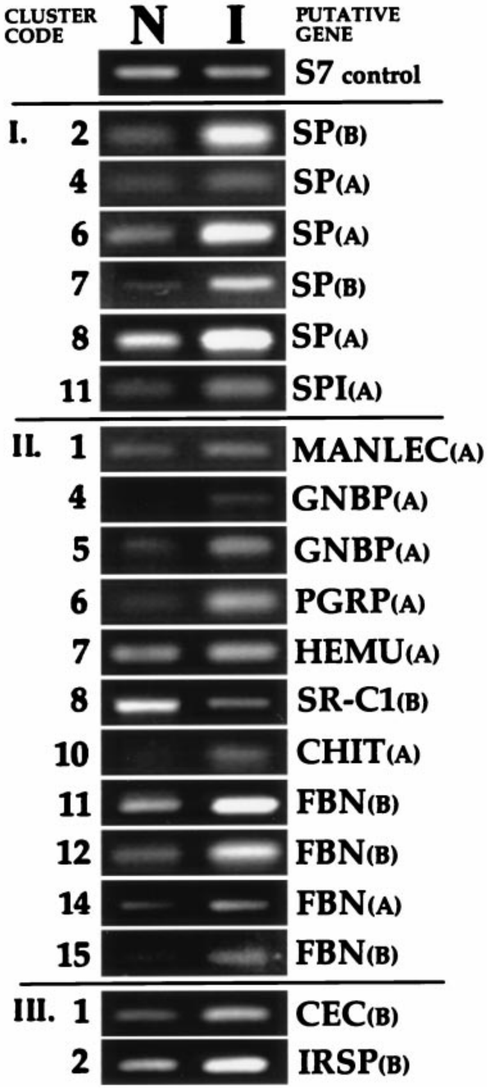

Fig. 2. Reverse transcription–PCR expression assays of putative immunity genes in naïve and bacterially challenged cell lines. Expression levels of the 38 *A. gambiae* putative immunity genes (Tables 2–4) were assayed by reverse transcription–PCR on RNA extracted from a naïve cell line (N) and a cell line that had been incubated with heat-killed *E. coli* and *Micrococcus luteus* for 8 h (I) as described (15). The cell line cDNA templates were normalized for the expression of the ribosomal protein S7 gene (S7), and the numbers of PCR cycles were empirically estimated for each transcript to avoid overamplification; specific primers were used for optimal amplification of products ranging in length from 250 to 500 bp at an annealing temperature of 58°C. Of the 38 putative immunity genes, one previously known was shown to be inducible but is not shown (II.3); 18 others, as shown, are transcriptionally activated in the cell lines 4A-3A (A) and 4A-3B (B). One putative receptor (II.8) is repressed by immune challenge.

Concluding Remarks. This pilot gene discovery project multiplied several fold the number of *A. gambiae* gene sequences that had been deposited in the public databases during a decade of increasing interest in the molecular genetics of this major malaria vector. The 8 previously identified and the 30 newly discovered putative immunity genes will contribute significantly to the dissection of *A. gambiae* innate immunity; their potential involvement in the mosquito's antiparasitic defense mechanisms is a matter to be addressed experimentally. The constantly increasing amount of expressed DNA that is sequenced from other organisms will permit identification of more homologues within the generated set of *A. gambiae* ESTs, thus increasing the number of mosquito genes with putative functions. Such an increase would be particularly useful for the 1,262 clone clusters that did not yield significant BLASTX hits to date. The normalized cDNA libraries that we have constructed from the cell lines remain a promising source for additional gene discovery through further large-scale EST determination. Similar, normalized libraries could be constructed from whole mosquitoes (or from isolated tissues such as midgut and salivary glands that are involved in the malaria life cycle), permitting discovery of important genes that may not be expressed significantly in the cell lines. Systematic expression analysis of the already available clone clusters with cDNA-microarray technology is expected to reveal additional immune-responsive components, developmentally regulated genes, or genes that may be induced by parasitic infections.

Note Added in Proof. The recent criticism of Wang *et al*. (47) that normalization/subtractive hybridization can lead to systematic loss of rare mRNA sequences bearing long poly(A) tails does not apply to the methods used in this study (23). These have been optimized to reduce the length of poly(A) tails to 26 ± 12 nucleotides as verified by sequencing; hybrids of that length do not get subtracted by HAP chromatography.

We thank Ryan Kinkaid, Vladan Miljovic, Greg Doonan, Robert Brown, Christy Smith, Jurgen Zimmermann, Barbara Schwager, Monica Benes, and Holger Erfle for sequencing; Brad Johnson and Keith Crouch for template production; M. Andrade for assistance with analysis of the sequences; A. Danielli, M. Lagueux, E. Levashina, and L. Moita for permission to refer to unpublished material; and H.-M. Müller for the hemocyte-like cell lines. This investigation received financial assistance from the United Nations Development Program/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases.

1. Collins, F. H. (1994) *Parasitol. Today* **10,** 370–371.
2. Curtis, C. F. (1994) *Parasitol. Today* **10,** 371–374.
3. Ribeiro, J. M., Nussenzveig, R. H. & Tortorella, G. (1994) *J. Med. Entomol.* **31,** 747–753.
4. Shen, Z., Dimopoulos, G., Kafatos, F. C. & Jacobs-Lorena, M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 6510–6515.
5. Arcà, B., Lombardo, F., de Lara Capurro Guimarães, M., della Torre, A., Dimopoulos, G., James, A. A. & Coluzzi, M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 1516–1521.
6. Dimopoulos, G., Richman, A., della Torre, A., Kafatos, F. C. & Louis, C. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 13066–13071.
7. Beier, J. C. (1998) *Annu. Rev. Entomol.* **43,** 519–543.
8. Collins, F. H., Sakai, R. K., Vernick, K. D., Paskewitz, S., Seeley, D. C., Miller, L. H., Collins, W. E., Campbell, C. C. & Gwadz, R. W. (1986) *Science* **234,** 607–610.
9. Vernick, K. D., Fujioka, H., Seeley, D. C., Tandler, B., Aikawa, M. & Miller, L. H. (1995) *Exp. Parasitol.* **80,** 583–595.
10. Dimopoulos, G., Richman, A., Müller, H.-M. & Kafatos, F. C. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11508–11513.
11. Dimopoulos, G., Seeley, D., Wolf, A. & Kafatos, F. C. (1998) *EMBO J.* **17,** 6115–6123.
12. Barillas-Mury, C., Charlesworth, A., Gross, I., Richman, A., Hoffman, J. A. & Kafatos, F. C. (1996) *EMBO J.* **15,** 4691–4701.
13. Zheng, L., Cornel, A. J., Wang, R., Erfle, H., Voss, H., Ansorge, W., Kafatos. F. C. & Collins, F. H. (1997) *Science* **276,** 425–428.
14. Richman, A., Dimopoulos, G., Seeley, D., Kafatos, F. C. (1997) *EMBO J.* **16,** 6114–6119.
15. Müller, H.-M., Dimopoulos, G., Blass, C. & Kafatos, F. C. (1999) *J. Biol. Chem.* **274,** 11727–11735.
16. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., *et al*. (1991) *Science* **252,** 1651–1656.
17. Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. M. (1999) *Nat. Genet.* **21,** Suppl., 10–14.
18. Lawson, D. (1999) *Parasitology* **118,** S15–S18.
19. Williams, S. A., The Filarial Genome Project, Johnston, D. A. & The Schistosome Genome Project (1999) *Parasitology* **118,** S19–S38.
20. Myler, P. J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickel, E., Sisk, E., Sunkin, S., *et al*. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2902–2906.
21. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., *et al*. (1998) *Science* **282,** 1126–1132.
22. Vizioli, J., Bulet, P., Lowenberger, C., Blass, C., Müller, H.-M., Dimopoulos, G., Hoffmann, J., Kafatos, F. C. & Richman, A. (2000) *Insect Mol. Biol.* **9,** 75–84.
23. Bonaldo, M. F., Lennon, G. & Soares, M. B. (1996) *Genome Res.* **6,** 791–806.
24. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
25. Bairoch, A. & Apweiler, R. (2000) *Nucleic Acids Res.* **28,** 45–48.
26. Tamames, J., Ouzounis, C., Casari, G., Sander, C. & Valencia, A. (1998) *Bioinformatics* **14,** 542–543.
27. Muta, T. & Iwanaga, S. (1996) *Curr. Opin. Immunol.* **8,** 41–47.
28. Paskewitz, S. M., Reese-Stardy, S. & Gorman, M. J. (1999) *Insect Mol. Biol.* **8,** 329–337.
29. Gorman, M., Andreeva, O. V. & Paskewitz, S. (2000) *Insect Biochem. Mol. Biol.* **30,** 35–46.
30. Danielli, A., Loukeris, T. G., Langueux, M., Müller, H.-M., Richman, A. & Kafatos, F. C. (2000) *Proc. Natl. Acad. Sci. USA* **97,** in press.
31. Söderhäll, K. & Cerenius, L. (1998) *Curr. Opin. Immunol.* **10,** 23–28.
32. Lathinen, U., Hellman, U., Wernstedt, C., Saraste, J. & Petterson, R. F. (1996) *J. Biol. Chem.* **271,** 4031–4037.
33. Ham, P. J. (1992) in *Advances in Disease Vector Research*, ed. Harris, K. (Springer, New York), pp. 101–149.
34. Yoshida, H., Kinoshita, K. & Ashida, M. (1996) *J. Biol. Chem.* **271,** 13854–13860.
35. Theopold, U., Samakovlis, C., Erdjument-Bromage, H., Dillon, N., Axelsson, B., Schmidt, O., Tempst, P. & Hultmark, D. (1996) *J. Biol. Chem.* **271,** 12708–12715.
36. Paerson, A., Lux, A. & Krieger, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 4056–4060.
37. Franc, N. C., Heitzler, P., Ezekowitz, R. A. & White, K. (1999) *Science* **284,** 1991–1994.
38. Lu, J. (1997) *BioEssays* **19,** 509–518.
39. Gokudan, S., Muta, T., Tsuda, R., Koori, K., Kawahara, T., Seki, N., Mizunoe, Y., Wai, S. N., Iwanaga, S. & Kawabata, S.-I. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10086–10091.
40. Hoffmann, J. A. & Reichhart, J.-M. (1997) *Trends Cell Biol.* **7,** 309–316.
41. Grosshans, J., Schnorrer, F. & Nusslein-Volhard, C. (1999) *Mech. Dev.* **81,** 127–138.
42. Ostrowski, J., Van Seuningen, I., Seger, R., Rauch, C. T., Sleath, P. R., McMullen, B. A. & Bomsztyk, K. (1994) *J. Biol. Chem.* **269,** 17626–17634.
43. Jiang, H., Wang, Y., Korochkina, S. E., Benes, H. & Kanost, M. R. (1997) *Insect Biochem. Mol. Biol.* **27,** 693–699.
44. Weiss, G., Wachter, H. & Fuchs, D. (1995) *Immunol. Today* **16,** 495–500.
45. Rouault, T. & Klausner, R. (1999) *Curr. Top. Cell. Regul.* **35,** 1–19.
46. Dunkov, B. C., Zhang, D., Choumarou, K., Winzerling, J. J. & Law, J. H. (1995) *Arch. Insect Biochem. Physiol.* **29,** 293–307.
47. Wang, S. M., Fears, S. C., Zhang, L., Chen, J.-J. & Rowley, J. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4162–4167.