European Molecular Biology Laboratory
University of Crete, School of Sciences, Department of Biology

Doctoral Dissertation

Functional Genomics and Developmental Transcriptomics of the mosquito malaria vector, *Anopheles gambiae*

Anastasios Koutsos

London, July 2006

European Molecular Biology Laboratory,
Meyerhofstrasse 1,
D-69117 Heidelberg,
Germany.

University Of Crete, Department of Biology,
Vassilika Vouton,
71409, Heraklion, Crete,
Greece.

Current Address:
Imperial College London,
Division of Cell and Mollecular Biology,
Sir Alexander Fleming Building,
Imperial College Road,
London, SW7 5TH,
United Kingdom.

Tel: + 44 (0) 207 594 5361
e-mail: a.koutsos@imperial.ac.uk

Cover picture: A photomosaic picture of a mosquito blood feeding on a human, composed of 1054 malaria related images. All images are from the TDR Malaria Image Library.

Functional genomics and developmental transcriptomics of the mosquito, malaria vector, *A. gambiae*

Anastasios Koutsos

PhD thesis

European Molecular Biology Laboratory
& University of Crete, School of Sciences, Department of Biology.

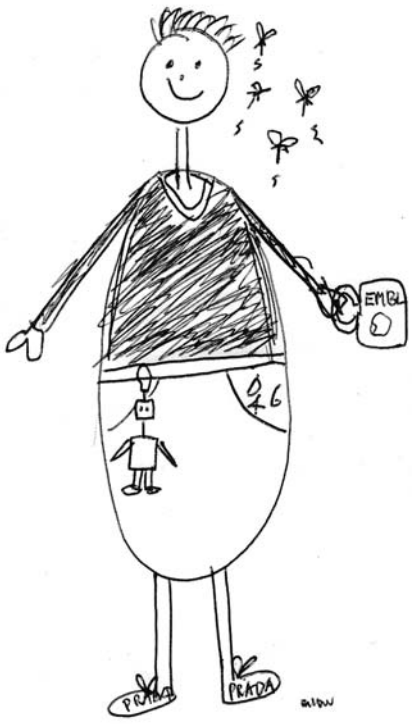
Supervisor:
Professor Fotis Kafatos

PhD Thesis Defense Committee:
Despoina Alexandraki, Associate professor, Un. of Crete, Department of Biology
George Christophides, Senior Lecturer, Imperial College London
Christos Delidakis, Associate Professor, Un. of Crete, Department of Biology
Matthias Hentze, Professor, European Molecular Biology Laboratory
Christos Louis, Professor, Un. of Crete, Department of Biology
Charalambos Savvakis , Professor, Un. of Crete, Department of Biology
Joseph Papamattheakis, Professor, Un. of Crete, Department of Biology
Nekarios Tavernarakis, PhD, Institute of Molecular Biology and Biotechnology

Dedicated to my family,
for making it all possible in the first place

Table of Contents

Chapter 4



Acknowledgements

Acknowledgements

The text you will read in the next 150 pages represents the outcome of 3.5 years of work. These pages were meticulously crafted in language in order to convey a precise and scientific message; they are the result of repeated editing and language correction: they come from the mind. The text that you will find here, however, aims to express gratitude to the people I have been collaborating with for those 3.5 years. These pages may be sloppy, informal and grammatically incorrect; they come straight from the heart.

The work presented in the next sections covers my joint PhD studentship in the European Molecular Biology Laboratory (EMBL), in Heidelberg, Germany and the University of Crete in Greece. The majority of the work was carried out at EMBL but the final year was carried out at Imperial College London, due to the lab's relocation. Since this dissertation represents the end of an academic period for me, I feel that I need to thank numerous people and I apologise for the people that I forget to mention here.

I wish to express my gratitude to Fotis C. Kafatos for agreeing to make me part of his scientific team. I will never know what made him choose me in the first place. I will also not pretend that everything went smoothly; this collaboration had its ups and downs. Nevertheless, I think that I need to thank him for giving me this great opportunity, allowing me to engage into additinal activities than research and, most importantly, teaching me one of the greatest lessons in politics.

I would also like to thank the other members of my Thesis Advisory Committee. First, to very warm thanks to Professor Matthias Hentze, who was my most immediate advisor and helped me in some of my difficult situations. Then, I would like to thank professor Kitsos Louis, Associate Professor Christos Delidakis, both whom I have known before from my studies at the University of Crete and Dr. Peer Bork, for being the other EMBL member in my TAC and giving me valuable advice. Special thanks to Rob Russel at EMBL, for help with the modeling of the LRR domain of LRIM1, presented in chapter 4.

I also wish to thank George K. Christophides for being my immediate supervisor and all the members of the Kafatos laboratory, past and present, who withstood my (sometimes annoying) presence in the lab. Especially, I would like to thank Stephan

Meister and Claudia Blass for forming together a small (microarray) team, at least in the first year of my thesis. However, there are also many more people that deserve a special mention.

First, I would like to thank all of Fotis' personal assistants. I had the pleasure to talk to them for many administration relevant (and sometimes irrelevant) matters and they were very helpful, especially in many urgent circumstances. So thanks very much Manuela, Natalie, Michael, Mehrnoosh, Andrea in Heidelberg and David and Rabeya in London.

Undoubtedly, I would also like to mention all those people I collaborated with at my other projects. To the Greek Mafia at EMBL and especially to Doros, for organizing the 'Greek party'. It all started from a somewhat vague and crazy idea, but soon appeared to be one of the funniest and enjoyable experiences that brought together most – if not all- of the Greeks.

Perhaps the best example of a project that highlighted to me the value of teamwork, rather than the pursuit of personal ambitions was the organising of the 5th International EMBL PhD students symposium "Design of Life, Learning from Nature". When we first started, I had my doubts that we would go all the way through and my suspicions were proven right in our initial, lengthy and chaotic meetings. Nevertheless, the organizing committee of this symposium not only found its pace but managed to deliver a great symposium. The funding of the symposium by a four-year European Commision Marie Curie grant has perhaps been considered by many as the best highlight. For me, it is clearly not. What I will always remember is that our committee worked without a 'central coordinator'. 'Teamwork can be achieved by a decentralized parameter' said one of our invited speakers and he couldn't have been more true. I, therefore, wish to thank all the people in this committee, as they themselves taught me that instead of crude logic, intensive (and somewhat constipated) planning, there is always a time to be a little bit more chaotic, more relaxed and maybe have some fun.

During the first year in Heidelberg I made some new friends: Maria, Dilem, Fabian, Andreas and Nadinne. Unfortunately, due to lack of time, or maybe other things, I lost news for some of them. I want to specifically thank Nadinne for 'putting me in the gang', for keeping in touch and for the fantastic discussion we had in a warm evening in Rome and Dilem, with whom I enjoyed our spin racing course in the gym. I would also like to thank my other immediate friends who have always

been there for me and mostly coped with my problems at work. To Alexandra, Theodora and Dimitris and our discussions about the old days of Crete, to Valia for learning me how to cook a traditional ‘greek pitta’, to Piyi and our hilarious moments of trying to repair an inflatable mattress and to Melpi for our common exquisite taste and our numerous visits to the shop in Heidelberg that sells designer furniture.

Undoubtedly, the move from Heidelberg to London marked the final year of my PhD. I thank my labmate Ellen Runn for drawing those funny sketches of myself that are included as special covers in my thesis chapters. A very special thanks is reserved for the people of Robert Sinden’s lab for being welcoming, and especially to Chandra, for bearing with solitude my endless talking. I would also like to thank Professor Robert Sinden, whose photographic memory of malaria related information still amazes me. He has been spending much of his precious time in listening to my crazy ideas of how parasite infection is mediated and giving me valuable advice. From the London times, I also like to thank my friend Lila, for keeping in touch with me all those years and Dimitris, for reasons that he knows himself.

One of the people from Heidelberg that I greatly miss is the person I was sharing the bench with; Rui. I miss our discussion about Chinese traditions, as well as my weekly “Chinese learning bit”. In addition, Dolores Doherty, our insectary technician, has been an oasis of relaxation amidst my stress and problematic period of working. I really much appreciated our talks.

Undoubtedly, however, the person that I miss most from Heidelberg times has been Stephanie. Apart from being a valuable friend she was the person I would always go to for advice. I still cherish our endless discussions about mosquitoes, Macs, biking, swimming, cooking and everything else. I was privileged with a gift of collaboration with her, even though it has been in the very final stages of my PhD. I wish to her even greater success, although I am confident that she will have it.

Finally, when we left Heidelberg nobody else managed to complement Stephanie’s void than Sofia. Sofia is always there to talk and to get advice. Every time I go to her with a theory of how parasite infection proceeds, she tells me that I am wrong! Then I get annoyed and tell her that I will prove her wrong! I have not yet managed to do it, though. I was also privileged to collaborate with her in an exciting project concerning microarrays.

Finally, as in my previous work, the last acknowledgements are always reserved for something unexpected. In this case, it is not that unexpected; I feel that the

decision was a long time ago. I am indebted to Alexandra Manaia and Julia-Willingale Theune, the Education officers of the European Learning Laboratory for the Life Sciences (ELLS) of EMBL, for helping me to creat the educational activity "The Virtual Microarray", which was based on one of my ideas. This activity was conceived in the first year of my PhD thesis and aimed to explain to high school teachers how scientists perform microarray experiments. The 'Virtual Microarray activity' ran almost parallel to my normal research activities. The excellent collaboration I had with them, the fun we had developing the activity and the time that we took in showcasing the activity in numerous educational meetings made me very enthusiastic about science education and marked the project as the best one I have ever been involved in my academic years. Naturally, I want to say a very warm thank you to them for being everything.

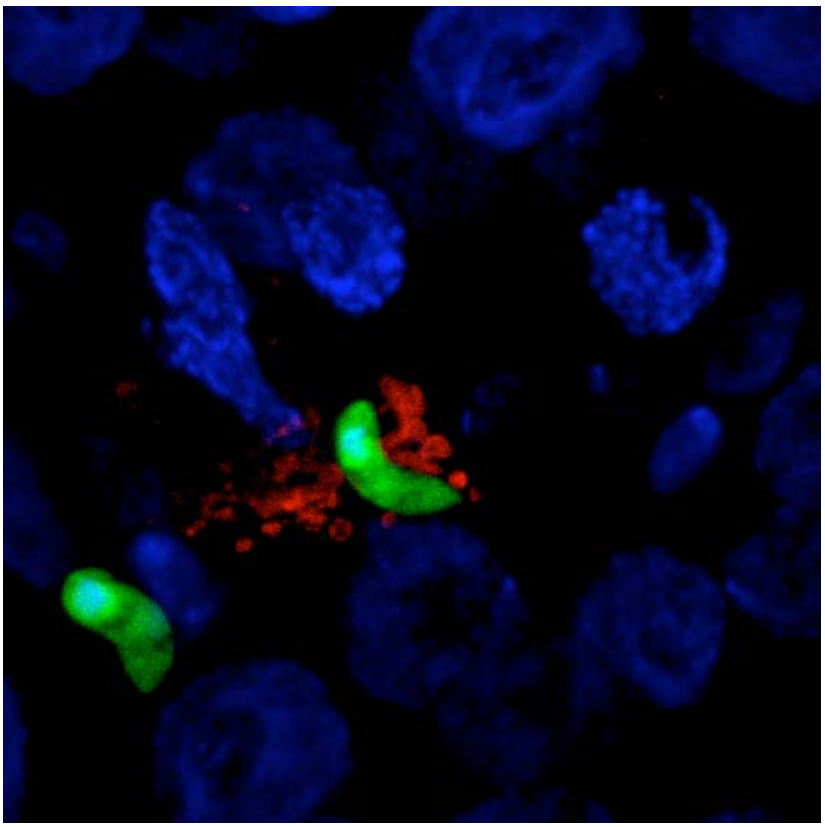
During these 3.5 years I spend a limited amount of time with my family. I thank them for being there, knowing that I can still not make the promise to see them more often.

In finishing this PhD thesis, many things come to mind. During those years, there have been many times of hard work, many unsuccessful experiments, many disasters, much distress but also periods of success and hapinness. 'Ithaca has given you the splendid voyage. Without her you would have never set out. But she has nothing more to give you. And if you find her poor, Ithaka has not deceived you. So wise that you became, with such experience, that already you will have understood, What these Ithakas mean to you' concludes my favourite Greek poem by Kavafis. For me, the meaning of these Ithakas lies hidden among the following pages.

London, 31st July 2006.

In 1730 Dr Thomas Fuller wrote:

‘Can any man, can all the Men in the World, tho’ assisted by Anatomy, Chymistry, and the best Glasses, pretend positively and certainly to tell us, what particles, how sized, figured, situated, mixed, moved, and how many of them, are requisite to produce a quartan ague, and how they specifically differ from those of a tertian … ?’



We are now able to tell all these things. They have been written in hundreds of books, and are familiar to thousands of students. Those who belittle the powers of science are not always, perhaps, the wisest of men.

The history of malaria contains a great lesson for humanity – that we should all be more scientific in our habits of thought, and more practical in our habits of government.

Sir Ronald Ross, 1910

Abstract

The mosquito *Anopheles gambiae* is the major vector of human malaria in Africa. Malaria is caused by parasites of the genus *Plasmodium* which undergo a complex developmental cycle inside the mosquito. These parasites are then transmitted to humans by infectious bites and cause the clinical manifestations of the disease. Although mosquitoes were identified as vectors of the disease several decades ago, little knowledge was available, especially at the molecular level. This situation was improved the last two decades largely due to increased interest by researchers, which led to the publication of the *A. gambiae* genome, alongside the *Plasmodium falciparum* genome and the development of specific tools for functional genomic studies.

The present study concentrates in post-genomic research of *A. gambiae* and describes efforts in several fields. In the field of bioinformatics, we used the genome information to position all publicly available expressed sequences (ESTs and cDNAs) to construct expressed contigs. This information was presented in a newly developed database, AnoEST and was supplemented with functional annotation information, which is valuable for the analysis of microarray experiments. In addition, this study provided evidence for the existence of several expressed sequences that have been missed by the automatic gene prediction pipeline of Ensembl.

In the field of transcriptomics, we constructed a new microarray platform, MMC1 that encompasses 20,000 ESTs from *A. gambiae*. We used this platform to monitor global gene expression in nine different time periods of the lifecycle of *Anopheles* and four different tissues of the adult mosquito. Our analysis identified developmental programmes and tissue-specific patterns and showed that genes which belong to related functional categories, or that encode the same or functionally linked protein domains are clustered together. Comparative analysis of our data together with published data from *Drosophila melanogaster,* which diverged from *Anopheles* some 250 million years ago, revealed high correlation of developmental expression between orthologous genes. The degree of gene expression similarity is not correlated with the degree of coding sequence similarity, implying uncoupled evolution of gene expression profiles and coding sequences. This is the first large-scale comparative

transcriptomic analysis in insects which detected important evolutionary features of insect transcriptomes.

In the field of functional genomics, we present a comprehensive functional survey of *leucine rich repeat immune gene 1*, *LRIM1*, and its relation to *Anopheles* innate immune responses. We showed that *LRIM1* is involved in responses against pathogenic bacteria and argue that the response is dependent on bacterial species and bacterial concentration. Finally, we demonstrated involvement of *LRIM1* in the killing and melanisation of the *Plasmodium berghei* malaria parasites and showed evidence recruitment and localisation in close proximity to the malaria parasites.

Thus, the multifaceted analysis presented in this thesis aims to highlight different aspects of *A. gambiae* research: bioinformatic and transcriptional studies that promote knowledge in mosquito basic biology and fuctional analyses that aim to identify important factors of the mosquito immune system. Our integrated approach in the study of *A. gambiae* may prove useful towards effective future vector control strategies against the malaria parasite.

Περίληψη

Το κουνούπι *Anopheles gambiae* είναι ο κύριος φορέας της ελονοσίας στην Αφρική. Η ελονοσία προκαλείται από παράσιτα του γένους *Plasmodium* τα οποία υφίστανται ένα πολύπλοκο αναπτυξιακό κύκλο στο κουνούπι. Αυτά τα παράσιτα μεταφέρονται αργότερα στους ανθρώπους με μολυσματικά τσιμπήματα και δημιουργούν την κλινική εικόνα της ασθένειας. Αν και τα κουνούπια αναγνωρίστηκαν ως φορείς της ελονοσίας πριν από πολλές δεκαετίες, λίγη γνώση υπήρχε διαθέσιμη για τους οργανισμούς αυτούς, ιδιαίτερα σε μοριακό επίπεδο. Η κατάσταση αυτή βελτιώθηκε τις δυο τελευταίες δεκαετίες κυρίως λόγω του αυξανόμενου ενδιαφέροντος των επιστημόνων, γεγονός που οδήγησε στη δημοσίευση του γονιδιώματος του *A. gambiae*, ταυτόχρονα με το γονιδίωμα του *Plasmodium faliparum* και στη δημιουργία συγκεκριμένων εργαλείων για λειτουργικές γονιδιωματικές μελέτες.

Η παρούσα εργασία εστιάζεται στη μεταγονιδιωματική έρευνα στο κουνούπι, *A. gambiae* και περιγράφει προσπάθειες σε πολλά πεδία. Στο πεδίο της βιοπληροφορικής, συλλέξαμε όλες τις εκφραζόμενες αλληλουχίες (ESTs και cDNAs) από βάσεις δεδομένων και χρησιμοποιήσαμε την πληροφορία του γονιδιώματος για να τις συστοιχίσουμε και να δημιουργήσουμε contigs. Η πληροφορία αυτή παρουσιάστηκε σε μια καινούργια βάση δεδομένων, την AnoEST. Σε αυτήν προστέθηκε λειτουργική πληροφορία για τα γονίδια, η οποία είναι πολύτιμη για την ανάλυση πειραμάτων μικροσυστοιχιών. Επιπροσθέτως, αυτή η μελέτη έδωσε αποδείξεις για την ύπαρξη πολλών εκφραζόμενων αλληλουχιών, οι οποίες είχαν αγνοηθεί από τον αλγόριθμο αυτόματου χαρακτηρισμού γονιδίων της βάσης δεδομένων Ensembl.

Στο πεδίο της μεταγραφής, δημιουργήσαμε μια καινούργια μικροσυστοιχία, την MMC1, η οποία περιλαμβάνει 20 000 ESTs από το κουνούπι *A. gambiae*. Χρησιμοποιήσαμε τη μικροσυστοιχία αυτή για να παρακολουθήσουμε τη συνολική γονιδιακή έκφραση σε 9 διαφορετικές περιόδους του κύκλου ζωής του *A. gambiae* και σε 4 διαφορετικούς ιστούς του ενήλικου κουνουπιού. Η ανάλυσή μας ταυτοποίησε αναπτυξιακά προγράμματα και ιστο-ειδικά πρότυπα και κατέδειξε ότι γονίδια που ανήκουν σε σχετιζόμενες λειτουργικές κατηγορίες, ή ότι κωδικοποιούν τα ίδια ή λειτουργικά σχετιζόμενα πρωτεϊνικά κομμάτια, ομαδοποιούνται μαζί.

Συγκριτική ανάλυση των δεδομένων μας με αντίστοιχα δημοσιευμένα δεδομένα από τη φρουτόμυγα, *Drosophila melanogaster*, η οποία διαχωρίστηκε εξελικτικά από το κουνούπι περίπου 250 εκ. χρόνια, αποκάλυψε υψηλή συσχέτιση αναπτυξιακής γονιδιακής έκφρασης σε ορθόλογα γονίδια. Το βαθμός συσχέτισης της γονιδιακής έκφρασης δε σχετίζεται με το βαθμό ομοιότητας των κωδικών περιοχών, γεγονός που υποδηλώνει μη σχετιζόμενη εξέλιξη στην γονιδιακή έκφραση και κωδική αλληλουχία. Αυτή είναι η πρώτη ευρεία συγκριτική ανάλυση γονιδιακής έκφρασης στα έντομα, η οποία αναγνώρισε σημαντικά εξελικτικά στοιχεία.

Στο πεδίο της λειτουργικής γονιδιωματικής ανάλυσης, παρουσιάζουμε μια λειτουργική μελέτη του γονιδίου, *leucine rich repeat immune gene, LRIM1* και τη σχέση του με το μηχανισμό εγγενούς ανοσίας του *Anopheles*. Σε αυτήν τη μελέτη δείξαμε ότι το *LRIM1* εμπλέκεται στην ανοσολογική απάντηση εναντίων βακτηρίων και διατυπώνουμε ότι η απάντηση αυτή σχετίζεται με τα είδη των βακτηρίων και τη συγκέντρωσή τους. Τέλος, δείχνουμε ανάμειξη του *LRIM1* στη θανάτωση και στον μελανωτικό εγκλεισμό του παράσιτου, *Plasmodium berghei* και παρουσιάζουμε αποδείξεις για τη στρατολόγηση και τον εντοπισμό της πρωτεΐνης σε στενή γειτνίαση με το παράσιτο.

Συμπερασματικά, η πολυπρόσωπη ανάλυση που παρουσιάζεται σε αυτήν την εργασία στοχεύει να καταδείξει διαφορετικές πλευρές της έρευνας του κουνουπιού *Anopheles gambiae*: μελέτες βιοπληροφορικής και μεταγραφής, οι οποίες προάγουν τη γνώση στη βασική βιολογία και λειτουργικές μελέτες που στοχεύουν στην αναγνώριση παραγόντων του ανοσοποιητικού συστήματος του κουνουπιού. Η ολοκληρωμένη προσέγγιση στη μελέτη του *A. gambiae* μπορεί να φανεί χρήσιμη σε μελλοντικές αποτελεσματικές στρατηγικές ελέγχου του φορέα σε μια προσπαθεια καταπολέμησης του παράσιτου της ελονοσίας.

List of abbreviations

Abbreviations are sorted in order of chapter appearance

Chapter 1

A. gambiae	*Anopheles gambiae*
D. melanogaster	*Drosophila melanogaster*
DALY	disability adjusted life years
DDT	dichlorodiphenyltrichloroethane
LRIM1	leucine rich repeat immune gene 1
P. falciparum, P. vivax, P. malariae, P. ovale	*Plasmodium falciparum, vivax, malariae, ovale*
p.f.	post infectious blood meal
RBC	red blood cell
s.l	*sensu lato*
s.s	*sensu stricto*
WHO	World Heath Organisation

Chapter 2

4K	four thousand
cDNA	complementary DNA
CPU	central processing unit
EST	Expressed Sequence Tags
GO	gene ontology
NCLAG	No uniquely matched cluster of *A. gambiae*
SR	standard reference
TCLAG	Transcribed cluster of *A. gambiae*
UCLAG	Unaligned cluster of *A. gambiae*

Chapter 3

CEC	cecropin
CLIP	chymotrypsin-like serine proteases
cmRNA	complementary mRNA
CTL	c-type lectin
DD	developmentally declining
DEF	defensin
DI	developmentally increasing
EH	embryo high
EO	embryo low
FH	female high
GAM	gambicin
GNBP	Gram-negative binding protein
LH	larva high
LLH	late larva

LO	larva low
MMC	mosquito microarray consortium
mya	million years ago
OBP	odorant binding protein
PGRP	peptidoglycan recognition protein
PH	pupa high
PPO	pro-phenoloxidase
SCRB	scavenger receptor
SOM	self-organising maps
SRPN	serine protease inhibitor
TEP	thioester containing protein

Chapter 4

aa	αmino acid(s)
Ab	antibody
dsGFP, dsLRIM, dsCTL4, dsCTLMA2	*Double stranded RNA of GFP, LRIM1, CTL4 and CTLMA2*
E. coli	*Escherichia coli*
GFP	green fluorescent protein
GPI	glycosyl phospahtidyl inositol
KD	knockdown
LPS	lipopolysaccharides
LRR	leucine rich repeat
MBS	m-Maleimidobenzoyl-N-hydroxysuccinimide ester
NBS	nucleotide binding
o/n	overnight
PBS	phosphate buffered saline
PGN	peptidoglycans
R	Refractory strain
RNAi	RNA interference
RT	room temperature
RT-PCR	real time PCR
S	Susceptible strain
S. aureus	*Staphylococcus aureus*
TLR	Toll like receptors
TM	transmembrane

Chapter 1

The malaria disease burden

General Introduction

Vector borne diseases have accounted for some of the most devastating health problems of humanity. The causal agents of such diseases are parasites or viruses. Parasites require almost exclusively vector species – typically insects – to complete stages of their complex developmental cycle and are usually transmitted to humans by infectious bites, causing the clinical manifestations of the disease. Examples of vector-borne diseases include malaria, African sleeping sickness (African trypanosomiasis), leishmaniasis, lymphatic filariasis, Chagas disease, and oncocerciasis. Among them, malaria has been the most devastating parasitic disease in the world, with over half of the global population being at high risk and approximately 300 million cases reported every year.

Historic perspective of malaria

Malaria is thought to have been present throughout the history of mankind. Manifestations of the disease have been evident in prehistoric times (Angel, 1966; Nozais, 2003). Hippocrates was the first to make a detailed clinical description and determine the prevalence of the disease in the Mediterranean world. Romans associated malaria to swamps and mosquitoes and thus devised special draining and drying out procedures for stagnant water. The disease was named malaria from the Latin word for bad air ("mal aria"), as it was believed to be caused by poisonous air from swamps. It was only in 1880, that Charles Louis Alphonse Laveran (1845-1922) discovered parasites in the blood of soldiers suffering from fever and put forward the hypothesis that these organisms had been responsible for it. Battista Grassi (1854-1925) and his colleagues, Amico Bignami and Guissepe Bastianelli, identified mosquitoes as the transmission vectors and proved that only certain species of the genus *Anopheles* were able to transmit the disease in humans. In 1895, Ronald Ross discovered oocysts – a specific developmental stage of malaria parasites – in the midgut wall of mosquitoes and then went on to determine the entire life cycle of the avian parasite. Both Laveran and Ross received the Nobel prize award in 1907 and 1902, respectively, for their pioneering work in malaria research.

Since then, it has been established that malaria is caused by a unicellular apicomplexan protozoan of the genus *Plasmodium* that requires two hosts to

complete its life cycle: an invertebrate host or vector, typically mosquitoes of the genus *Anopheles* and a vertebrate host, humans or other vertebrates. In humans, malaria is caused by 4 different parasite species: *Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae*; among them, *P. falciparum* is the most dangerous and responsible for many deaths in humans throughout history.

Life cycle of the human malaria parasite

The lifecycle of the *Plasmodium* parasites can be understood as a sequence of four phases: one sexual in the mosquito vector and three asexual, one of which occurs in the mosquito vector and two in the human host (reviewed in (Ghosh et al., 2000; Knell and Wellcome Tropical Institute., 1991; Sinden, 1999)). The complete lifecycle of *Plasmodium* parasites is depicted in Fig. 1.1, although some differences may occur between *Plasmodium* species.

The blood feeding of a female mosquito on an infected human can be considered the start of the sexual phase of the parasite lifecycle (Fig. 1.1, lower middle). The mosquito ingests a substantial blood meal containing a small amount of *Plasmodium* G_0-arrested gametocytes whereas the remaining gametocytes die within the peripheral blood of the vertebrate host. In the mosquito vector, gametocytes escape from red blood cells within minutes to undergo differentiation and form the female and the male gametes. The process of gametogenesis has been studied in detail and at least two factors are thought to be essential: a sudden drop of temperature by at least 5° C and the presence of xanthurenic acid in the mosquito (Billker et al., 1998). The male gamete undergoes further differentiation resulting in the sudden release of 8 flagella in a process termed ‘exflagellation’. Each flagellum tears off and produces a mature male spermatozoon, which swims freely until it encounters a female gamete. Fertilisation by nuclear fusion ensues (30 min post infectious blood meal, p.f.) and produces a zygote, which then transforms into a motile, ‘banana-shaped’ structure called ookinete (9-24h pf). The ookinete is the invasive stage of the parasite that crosses two barriers in the mosquito: the peritrophic membrane - an acellular chitinous mesh approximately 2-10 μm in thickness, synthesized shortly after ingestion of the blood meal, which coalesces around the entire bolus of remaining erythrocytes (Shao et al., 2001)- and the midgut epithelium. After crossing those

barriers, the ookinete rests in the space between the epithelium and basal lamina and develops into an oocyst.

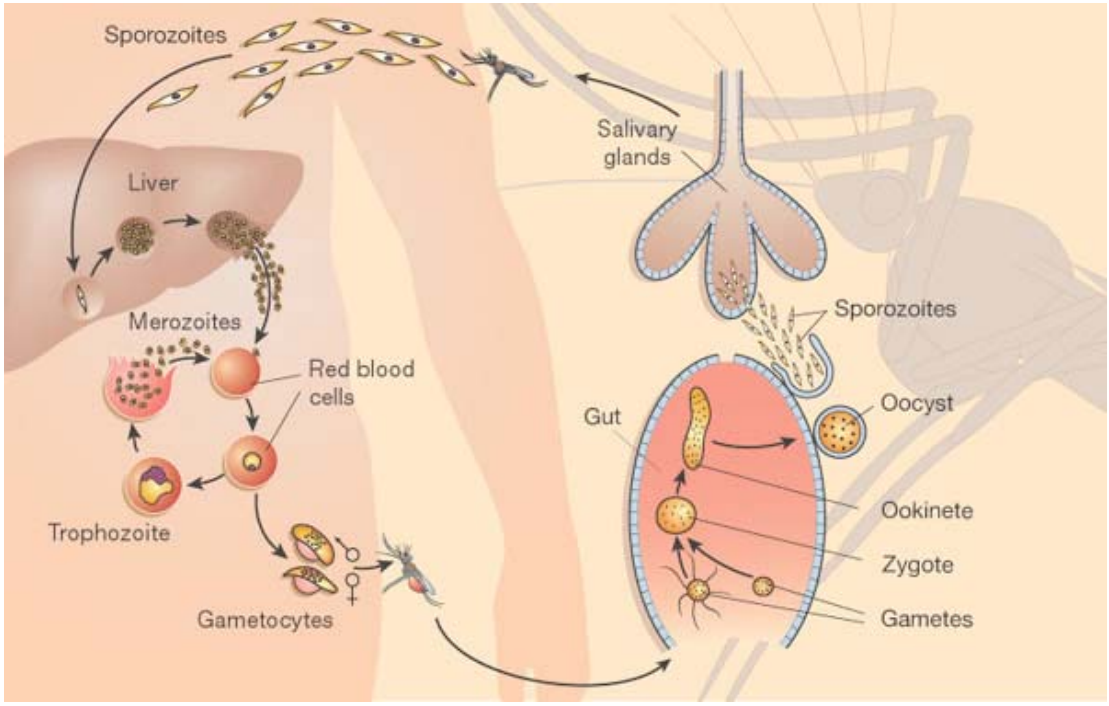


Figure 1.1. The *Plasmodium* life cycle. Schematic representation of the different parasite stages in the host (left) and the mosquito vector (right). (From (Wirth, 2002))

Secretion of parasite-derived chitinases dissolves the peritrophic membrane locally (Langer and Vinetz, 2001), allowing the parasite to pass to the ectoperitrophic space. For midgut invasion, it remains controversial whether the ookinetes migrate inter- or intracellularly as different routes have been reported for various mosquito-parasite combinations (Han et al., 2000; Meis et al., 1992; Shahabuddin and Pimenta, 1998; Vlachou et al., 2004; Zieler and Dvorak, 2000). Recently, a new model has been proposed (Baton and Ranford-Cartwright, 2005), which argues that ookinete entry is always intracellular, but the parasite route can either be intracellular or extracellular. However, it is not known if this model applies to all vector combinations or even all ookinetes within a particular combination.

The first asexual phase of parasite development starts with the oocyst undergoing a complex maturation period that typically lasts 2 weeks and culminates with the release of thousands of sporozoites in the mosquito haemocoele. The sporozoites swim freely in the haemocoele until they encounter the salivary gland epithelium, which they actively invade using gliding motility (reviewed in (Kappe et al., 2003)). Upon invasion, the sporozoites persist for many weeks in the extracellular space of

the gland and can be transmitted to a new host with the saliva upon a subsequent infectious bite. The gametocyte to sporozoite transition typically lasts three weeks, although it varies according to temperature and *Plasmodium* species.

The asexual phase in the vertebrate host (the second asexual phase of the cycle) begins with the bite of an infected female mosquito and the active 'crawling' of the sporozoites to the veins (Amino et al., 2005; Amino et al., 2006). The number of parasites transmitted by a single mosquito bite is not well established; however, as few as 2 to 10 sporozoites can initiate malaria infection in the field (Khusmith et al., 1994). Within minutes of entering the host's blood, they migrate to the liver, invade cells and transform to hepatic trophozoites (reviewed in (Frevert, 2004)). With the help of nutrients and temperature, the trophozoites can grow very rapidly, distending and destroying the hepatic cells. Later, they begin to multiply internally to form hepatic schizonts. One or two days after the infectious bite depending on the species of parasite, the hepatic cells are ruptured releasing numerous mature schizonts into the hepatic capillary, which subsequently invade erythrocytes within minutes of their release.

The final asexual stage of the parasite lifecycle takes place in the erythrocyte. A merozoite is released from the hepatic schizonts, enters the blood circulation and invades a red blood cell (RBC) and becomes the erythrocytic trophozoite, which ingests haemoglobin and acquires a characteristic malaria pigment. For two or three days, depending on the *Plasmodium* species, the trophozoites divide to produce new schizonts. The parasitized RBCs then erupt and release 8-16 merozoites each, which invade into new RBCs to start a new cycle. After several blood cycles, a proportion of trophozoites will develop via an alternative route to gametocytes and will continue to differentiate for several more days. The gametocytes remain in the cell membrane of the host RBCs and only invade the mosquito midgut after a subsequent infectious blood meal, which will reinitiate the *Plasmodium* lifecycle. The *Plasmodium* parasite is haploid during the majority of the lifecycle; the only diploid period is during the zygote to ookinete stages in the sexual phase in the mosquito. This is the reason that the mosquito is considered the parasite's definitive host.

The last phase of the parasite erythrocytic development causes the well-known disease manifestations in humans. Fever is induced when the schizonts rupture and trigger an immune response of the host which results in the release of pyrogens in the blood. The fever is irregular for 1-2 days and may remain irregular in severe malaria

caused by *P. falciparum*. The earlier hepatic stages in the human host do not produce any symptoms, as the parasites are rapidly cleared from blood circulation, possibly to avoid destruction from proteases in the blood. In *Plasmodium vivax,* some sporozoites do not immediately develop to trophozoites but become small dormant parasites called hypnozoites. Those hypnozoites may persist months, or even years, and may later begin development and cause relapsing cases of malaria. Finally, infected erythrocytes can rise up to 30 percent in *P. falciparum* and in other species, the malaria parasites can synchronise their cycle causing all schizonts to rupture simultaneously.

The malaria disease burden

It has been difficult to accurately determine the impact of malaria in global health; recent studies have estimated that malaria is affecting 300 million people annually, leading to approximately 2 million deaths (Breman, 2001) and 44 million disability adjusted life years (DALYs) – a quantitative factor that reflects the total amount of healthy life lost due to all causes of the disease. The disease exerts its heavy toll into young children and pregnant women, which together represent the two main risk groups. Malaria is endemic in more than 100 nations worldwide (Figure 1.2A). It has been estimated that the economic burden is high, accounting for a reduction of 1.3% in the annual economic growth rate and that the long term effect is a reduction to the gross national product of more that 50% (Sachs and Malaney, 2002). The fact that poverty (Fig 1.2B) is concentrated in the same geographical boundaries in the tropical and subtropical zones suggests that it is closely related. Indeed, the association of malaria with poverty in these areas is striking (compare Fig. 1.2A and 1.2B); malaria is responsible for major economical losses and, in retrospect, poverty results in increase in the disease burden.

The focus on mosquito vectors of malaria

Species of mosquitoes differ intrinsically in their vectorial capacity, which is their ability to sustain parasite development. Only approximately 30% of the *Anopheles*

spp transmit malaria and different mosquito species are the predominant vectors of malaria in different areas ((Kiszewski et al., 2004) and Fig. 1.3). In sub-Saharan Africa, where 90% of the cases of malaria occur, the predominant vectors are mosquitoes of the *Anopheles gambiae sensu lato (s.l.)* complex and *A. funestus.* The *A. gambiae s.l* complex includes 6 different subspecies: *A. gambiae sensu stricto (s.s.)*[1], *A. arabiensis, A. quadriannulatus, A. bwambae, A. merus and A. melas*, with the sister taxa relationships between this complex not clearly defined. The success of *Anopheles* vectors to transmit human pathogens results from the obligatory and repeated blood feeding of the adult females for egg production in combination with their extreme anthropophilic behaviour. They are found around human habitations, they are long lived and blood feed almost exclusively on humans. Mosquito longevity is especially important, as development of the parasite from gametocytes to infectious sporozoites can take many days depending on temperature. In addition, *A. gambiae* has the highest rate of *P. falciparum* sporozoite development. Thus, species of the *A. gambiae* complex are very efficient vectors for parasite transmission in sub-Saharan Africa and efforts to understand the disease burden in this area have largely coincided with population studies of *A. gambiae* as the principal disease vector.

[1] Unless otherwise noted, the name *A. gambiae* is used instead of *A. gambiae s.s* throughout the entire thesis.

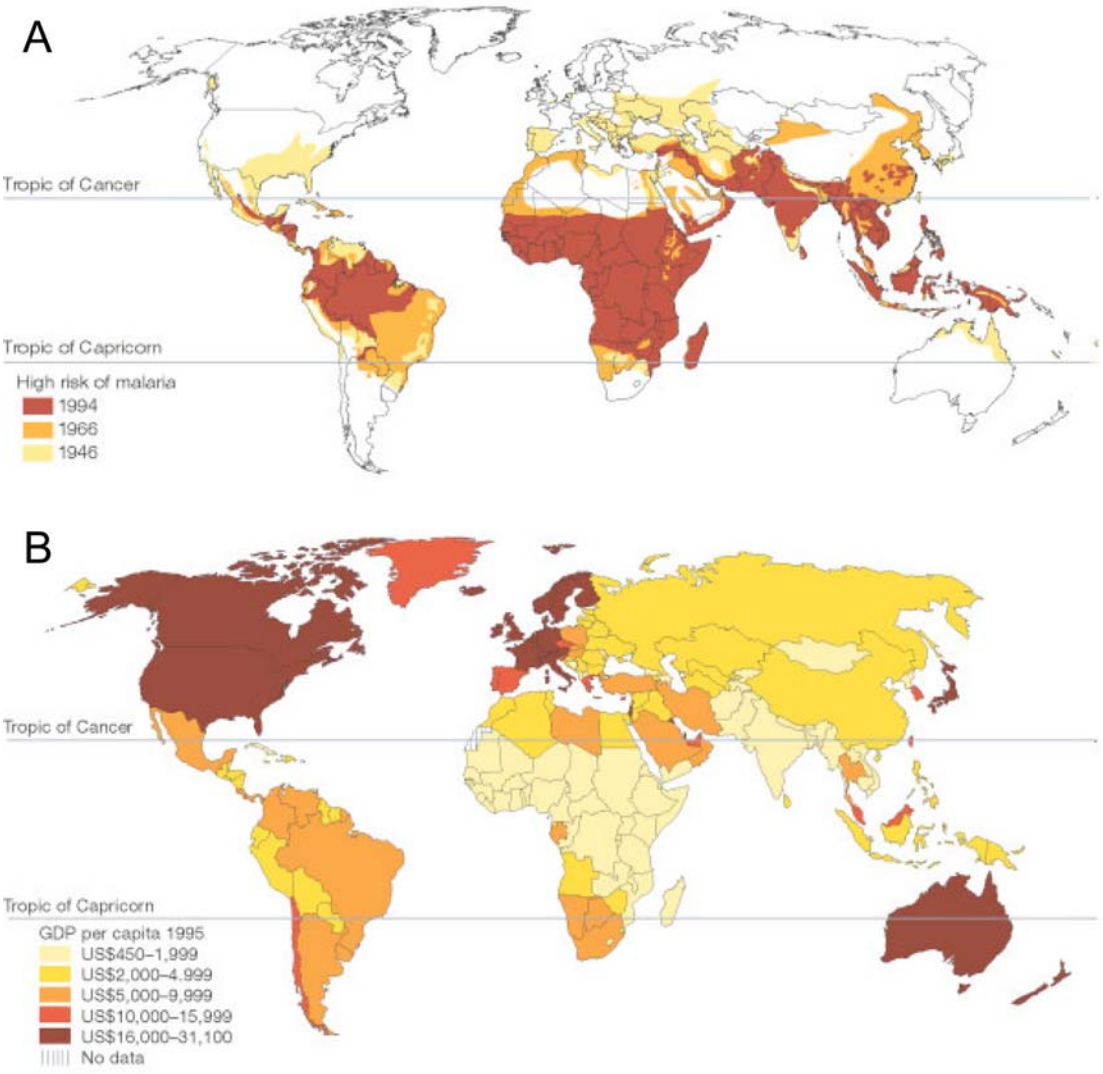


Figure 1.2. Correlation of malaria with poverty. A) Global distribution of malaria risk from 1946 to 1994. B) Global distribution of per capita gross domestic product (GDP). (Picture adapted from (Sachs and Malaney, 2002)).

Attempts for global malaria eradication in the 1960s

During the early 50-60s, the World Health organisation (WHO) launched a strategic action plan for global malaria eradication (reviewed in (Collins and Paskewitz, 1995)). Successful programmes had already been implemented in the United States, Europe and North America (compare malaria burden areas in 1946 and 1966 in Fig. 1.2A). The success of those programmes was due to the fact that they concentrated on several factors affecting malaria transmission and vector populations. Changes in agricultural uses of land, drainage of swamps, better housing and screening, education and the large-scale use of dichlorodiphenyltrichloroethane (DDT) were efficiently used to control mosquito populations. The goal of the WHO

initiative was similar: to interrupt parasite transmission by reducing, rather than eradicating, infected vector populations with the use of DDT. However, while eradication was most successful to temperate zones and isolated islands, the success has been short lived. Malaria re-emerged in areas where it had previously disappeared and mosquito resistance to DDT, which was initially observed in 1940 at low frequencies, became a common characteristic in many malaria endemic countries.

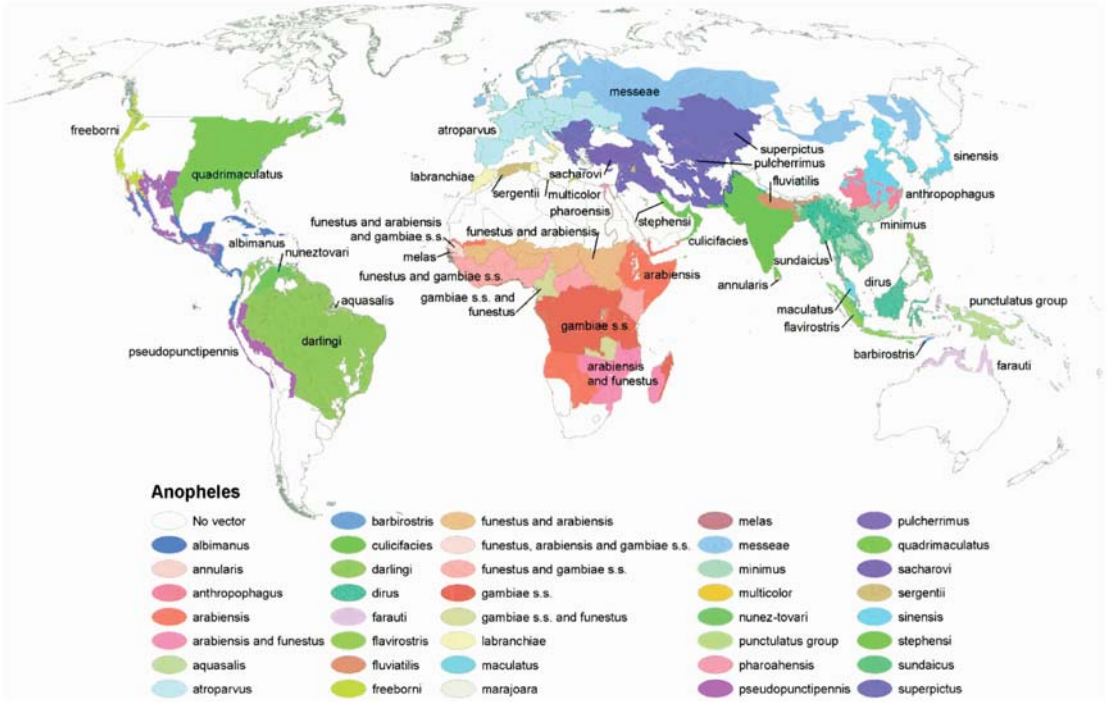


Figure 1.3. Global distribution of dominant or potentially important malaria vectors (from (Kiszewski et al., 2004))

The malaria situation has worsened in Africa during the past decades due to many other factors (reviewed in (Greenwood and Mutabingwa, 2002)); a series of environmental and climatic changes – global warming, floods associated with rains, construction of small dams - increased the number of local mosquito populations leading to an increase in infective bites. Political instability such as civil wars and refugee populations allowed the transfer of malaria from country to country. In addition, the collapse of health services and the increase in population is expected to double the number of people at risk. Undeniably, the factors that have mostly contributed to a dramatic increase in malaria are resistance of mosquito vectors to insecticides and resistance of *Plasmodium parasites* to drugs.

The majority of insecticides that have been used for vector control are common chemicals developed primarily for agriculture management. The first example of resistance appeared in the most commonly used insecticide, DDT and since then resistance to all other commonly used compounds such as carbamates, organophosphates and pyrethroids has been observed (reviewed in (Hemingway and Ranson, 2000)). The molecular basis of insecticide resistance is either the existence of mutations in target site genes or metabolic alterations of the level or activities of detoxification proteins (Hemingway et al., 2004). New compounds are not forthcoming, as manufacturers are reluctant to invest towards the development of insecticides for public health. Indeed the cost of the use of those compounds is prohibitive for use most of the African countries, representing an insufficient market for an investment.

Several drugs have been manufactured for the treatment of malaria symptoms. Earlier remedies included the administration of quinine, a substance extracted from the bark of chinchona trees and since then a number of synthetic drugs have been developed (reviewed in (Ridley, 2002)). However, compounds such as artemisinins, quinoline, antifolates and atovaquone/proguanil have been limited in their use due to high costs, rapid resistance and poor results. Chloroquine and antifolate sulphadoxine/pyrimethamine have been the only successful drugs to alleviate malaria symptoms. The very slow rate of parasite resistance to choloquine made it an ideal choice for large-scale use. However, with the current world-wide emergence of resistance to chloroquine and the developing resistance to sulphadoxine/ pyrimethamine, new drugs are urgently needed.

Humans are also displaying different degrees of susceptibility to malaria infections and several human factors have been proposed to affect parasite development in the blood stages. The majority of those cases deal with the absence of proteins or with abnormalities in the structure of RBCs, rendering the erythrocytes non permissive to parasite entry. Among them, sickle cell anaemia is a characteristic and well known example of a disease that alters the structure of RBCs upon oxygenation to produce sickle-shaped RBCs. Heterozygotes for this recessive autosomal trait have few number of sickle cells in their blood, which are not sufficient to cause the severe manifestation of the disease as in homozygotes but confer protection against malaria infections. Even if the person is infected, part of the RBCs cannot be infected, leading to decreased disease transmission. Due to this phenomenon of ‘heterozygote

advantage', the gene for sickle cell anaemia has been maintained in high frequencies in African populations. Other abnormalities include a and b thalassemias, glucose-6-phosphate dehydrogenase deficiency, Lewis and Kid Is (a) red cell type mutations and hereditary ovalocytosis. Several other molecules have been implicated with decreased malaria infection, such as the major histocombatibility molecule HLA-B53, the *Duffy* blood factors and the ICAM-1 putative receptor. Better understanding of the molecules that are necessary for parasite invasion of RBCs may result in the development of improved treatments and prevention of malaria.

New efforts to fight an old disease

The past decades have witnessed a deterioration of the situation of malaria in Africa. Malaria cases in epidemic counties and subsequently malaria-related mortality in sub-Saharan Africa are increasing. Also, *P. vivax* infections are reappearing in some areas, and resistance of mosquitoes to insecticides and/or parasites to drugs is emerging. To date a poor system for early diagnosis of the disease exists and no successful vaccine against malaria has been developed. This alarming situation prompted WHO to launch the "Roll Back Malaria" programme in 1998 with the aim to control the disease by the year 2010. This programme is a collaborative effort of international organisations, scientists, public health officials and African governments with the aim to limit the number of infectious bites to humans by a combination of different interventions and to deliver safe, cost–effective drugs for the treatment of malaria cases. In the past, successful methods implemented in one area have proven insufficient in other areas (Beier, 1998). Thus, with the benefit of hindsight, no individual method is likely to be the 'silver bullet' towards malaria control and the current approach aims to combine classical, successful methods with novel intervention strategies steming from malaria research. To this end, the sequencing of the draft genomes of both the major malaria vector, *A. gambiae* (Holt et al., 2002) and the most important parasite, *P. falciparum* (Gardner et al., 2002) has opened researchers unprecedented opportunities to envisage such strategies (Hoffman et al., 2002).

Only 3 out of 1,223 new drugs developed from 1975-1996 were antimalarials. The disappointing rate of new malaria drug development reflects the loss of interest by the industry. The parasite genome has already been used to underpin most of the

molecular advances in the search for novel drug agents and to investigate the molecular etiology to drug resistance. For example, scientists used homology information of parasite genes to plants and algae to identify several lipid biosynthesis genes, which are candidate targets for pharmaceutical approaches (Waller et al., 1998). Drug discovery is thus moving from molecular assays to cellular assays and proper animal models. Furthermore, an effective malaria vaccine that protects humans from the pre-erythrocytic stages of the disease has been the goal of researchers for many years and studies of the molecular properties of parasite cell invasion are expected to identify novel candidates for vaccine development.

The genome of *Anopheles* has also provided opportunities to develop strategies directed toward the mosquito vectors. These include the molecular investigation of insecticide resistance and the search for candidates for vaccine development. A more rational approach to the use of insecticides may increase their efficiency and prevent the rapid appearance of resistance; these methods include the use of insecticide mixtures and the adoption of a mosaic rotation strategy for their application. In addition, in the past years, stable transformation in mosquito vectors has been established (Catteruccia et al., 2000; Grossman et al., 2001). As a result, a novel strategy has been envisaged to produce genetically modified mosquitoes expressing factors that render them resistant to parasite development. Ongoing research is focusing on the search for mosquito factors that act as agonists or antagonists of parasite development (described also in chapter 4) and on the drive mechanisms that will establish them in wild mosquito populations. Whether this strategy will be transferred from the laboratory to the field and will be successful remains to be elucidated. Potential risks and ethical issues for the large-scale release of genetically modified mosquitoes will need to be assessed.

General aims of the current study

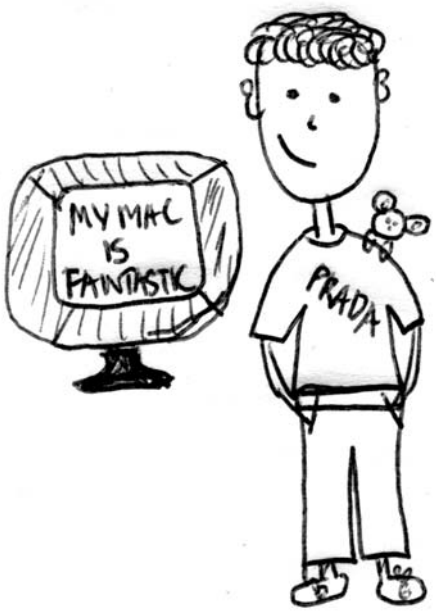
In conclusion, the current period of malaria research is governed by the impact of the genomics revolution in both vector and parasite. Investigation to the development of sustainable malaria control measures has turned from the field into the laboratory. Scientific research will unavoidably lead to an in-depth investigation of the disease transmission and basic biology of mosquitoes. More importantly, a thorough understanding of the interactions between the malaria parasite and the mosquito

vector at the molecular level is now possible. Such findings can then be tested in the field to assess their appropriateness to wild mosquito populations.

The work that has been carried out in this thesis was initiated in 2003, shortly after the publication of the mosquito genome sequence and aimed to develop an integrated approach towards functional genomic analysis in *A. gambiae.*

Chapter 2 describes the collaborative effort of researchers in our laboratory to utilise genomic information of expressed sequences to build a database called AnoEST. This database provided evidence for the existence of genes that were missed by the automated prediction pipelines and improved finding algorithms. Furthermore, the information of AnoEST served as the annotation platform for the microarrays, which were developed in our laboratory, as described in chapter 3. These microarrays were used to monitor global gene expression during 9 different developmental time periods in the mosquito lifecycle and 4 different tissues in adult mosquitoes. This study determined temporal and spatial specific gene transcripts and highlighted differences in the developmental processes between the mosquito *A. gambiae* and *Drosophila melanogaster*. It is the first large-scale comparative transcriptomic analysis in insects that provided insights to the evolution of coding sequences and orthologous gene expression. Finally, the thesis describes in the functional characterisation of LRIM1 a protein that is essential for mosquito antiparasitic defences. Results and further hypotheses on the mechanism by which LRIM1 it affects parasite development are discussed in chapter 4.

Altogether, the bioinformatics approaches for the construction of the AnoEST database, the transcriptomic analysis of mosquito development and the functional characterisation of the leucine-rich repeat family in *Anopheles* represent advances in both basic and vector-parasite biology of *Anopheles gambiae*. The combination of bioinformatic tools, high-throughoutput genome approaches and detailed molecular studies represent an integrated approach in malaria research and exemplify the efforts for the construction of specific tools and the production of in-depth knowledge, which are valuable towards future effective malaria control strategies.



Chapter 2

AnoEST: a genomic database for *Anopheles gambiae* functional studies

Introduction

The A. gambiae *genomic and bioinformatic tools for functional studies*

Completion of the draft sequence of the *A. gambiae* genome by an international scientific consortium in 2002 (Holt et al., 2002) provided researchers a vast array of opportunities for gene functional studies in the most important vector of human malaria. The initial gene prediction and annotation was a joint effort between Celera genomics and Ensembl (Birney et al., 2004). Since then, Ensembl frequently provides genome updates (Mongin et al., 2004) largely by using automated gene prediction algorithms and by incorporating manually curated gene models which have been contributed by individuals researchers. The release of the draft sequence of the genome also marked the release of AnoBase (Koutsos, 2002; Topalis et al., 2005), a database capturing genomic and biological information about *Anopheline* species with a specific focus on *A. gambiae*. AnoBase was created as a successor to an earlier database, AnoDB, and was based on the design of FlyBase (Drysdale and Crosby, 2005), the relational database of the *Drosophilidae*. The release of the genome sequence has thus provided researchers with a source for the construction of specific databases, capturing specialised information about *Anopheles*, which can be used for various gene functional approaches.

One of these approaches for the simultaneous assay of the activity of thousands of genes is DNA microarray profiling. These were initiated in *A. gambiae* with the sequencing of Expressed Sequence Tags (ESTs) prepared from cultured cells (Dimopoulos et al., 2000). Four thousand of these ESTs were used to construct the first mosquito cDNA microarray, the 4K-microarray platform (Dimopoulos et al., 2002). In addition, two other EST libraries were constructed from pooled developmental stages of *A. gambiae* (NAP1) or adult heads (NAH), and clones from these libraries have been sequenced. Twenty thousand of ESTs (from the NAP1, NAH and the 4K microarray libraries) were used to build a new cDNA microarray platform (MMC1 or 20K). MMC1 has been used for the transcriptomic profiling of the life cycle and of adult tissues of *A. gambiae* (discussed in more detail in chapter 3). However, insufficient annotation of the EST sequences hindered microarray

studies, and greatly limited the capacity of researchers to derive appropriate interpretations.

In the context of the *Anopheles* genome project, nearly 83,000 ESTs from naive and blood-fed adult mosquitoes were sequenced (Holt et al., 2002), and *in silico* analysis of these data detected upregulated genes in mosquitoes after a blood meal (Ribeiro et al., 2004). Furthermore, nearly 63,000 single reads from a full-length cDNA library were deposited in nucleotide databases by Genoscope (http://www.genoscope.org/). Today, over 200,000 EST or cDNA sequences are deposited in public sequence databases. This wealth of information about expressed sequences has provided the opportunity for the development of computation approaches to provide functional annotation to the 4K and MMC1 microarrays platforms, to facilitate the interpretation of the derived data.

Aims of the current study

DNA microarrays are useful tools for functional studies only if they are supplemented by rigorous and comprehensive gene annotation information. The focus of the current study is to provide the annotation platform for the newly constructed MMC1 microarray, which is based on approximately 20000 EST sequences of *A. gambiae*. Using the sequence information of the MMC1 ESTs along with any other publicly available expressed sequence, we have developed a pipeline for clustering expressed sequences. This method utilises genomic sequence as a template to map EST sequences to the genome and assemble overlapping ESTs to EST contigs. The contigs are then supplemented with automatic annotation information necessary for microarray functional genomic studies.

Methods

EST clustering

The analysis began with the collection and processing of all available *A. gambiae* EST and cDNA sequences, linked with their GenBank/EMBL-Bank/DDBJ accession number, clone name identifier, cDNA strand information and nucleotide sequence. All sequences were then aligned to the unmasked reference genome using the BLAT algorithm (Kent, 2002), considering all matches of 60 or more nucleotides, with at least 96% identity, a level that allowed for inaccuracies of EST sequences and polymorphisms and that captured all possible cross-hybridisations on a DNA microarray (Hughes et al., 2001). The accuracy of this step could be further improved at the expense of using orders of magnitude slower algorithms like sim4 (Florea et al., 1998).

ESTs were then assigned into groups (clustered) on the basis of their genomic overlap. For example, two sequences were assigned to the same contig if their overlap over the aligned regions (exons) was greater than a certain threshold (30nt in the current version of AnoEST). To avoid CPU-consuming all-against-all EST comparisons, which would be computationally challenging when considering potential alignment of over 200,000 EST sequences with nearly 500,000 genomic loci, we compared ESTs only with the contig's projection on the genome. DNA strands were considered independently. EST sequences originating from the 3′-end of a clone were deposited in public repositories as reverse complements; therefore we altered their alignment strand information prior to clustering.

In many cases, an expressed sequence could be aligned to more than one place in the genome (as it is the case for paralogues, transposable elements e.t.c.), making it difficult to reliably identify which genomic locus is actually represented by the EST. To address this we ranked EST to genome alignments using a number-of-matches minus number-of-mismatches scoring scheme, similar to BLAT. The matches with highest score were then marked as 'best', or as 'unique best' when the second-best score was significantly lower (e.g. by more than 15, to reflect the EST sequence error

rate and weak support from the data distribution). Contigs[2] including at least one 'unique best' EST were identified as TCLAG (for Transcribed CLuster of *Anopheles Gambiae*, also referred to as T-contigs below), whereas those sharing regions of high sequence identity to EST/cDNA sequences but there was no one sequence aligned to the locus as 'unique best' are identified as NCLAG contigs (with No uniquely matched ESTs). The third type of contig identifiers, UCLAG, corresponded to ESTs that failed to align (Unaligned) to the *A. gambiae* nuclear or mitochondrial genome. In the final step of the clustering procedure, we joined contigs that contained ESTs originating from the 5′- and 3′-ends of the same clone, provided that they mapped as 'unique best' to the corresponding EST contigs and were on the same chromosome, the same strand and less than 30kb apart.

The choice of many of the described parameters reflected a conservative approach that attempted to minimise errors of joining independent expressed loci at the expense of allowing some fragmentation errors, e.g. one gene could be represented by two EST contigs if there was no sufficient information to link these contigs together. The observed representation of 10,726 Ensembl gene models by 11,608 T-contigs suggested only a minor number of fragmentation artefacts. Use of strand specific clustering avoided the severe problems of erroneous joining of distinct genes (data not shown). However, some sequences that were inserted into the plasmid vectors in the wrong orientation could form erroneous contigs on the strand opposite the actual genes. An upper estimate of such errors is about 11%, counting the number of T-contigs overlapping annotated genes with respect to T-contigs on the opposite strand without annotation (counting overlaps over an average of 70%).

Automatic annotation

The derived contigs of expressed sequences were identified with gene models predicted by the Ensembl annotation pipeline, noting the fraction of genomic overlap over all predicted exons and allowing +/-150nt to capture EST contigs derived from UTRs.

[2] EST contigs have been initially named EST clusters in the publication of AnoEST Kriventseva, E. V., Koutsos, A. C., Blass, C., Kafatos, F. C., Christophides, G. K., and Zdobnov, E. M. (2005). AnoEST: toward A. gambiae functional genomics. Genome Res *15*, 893-899.. To avoid confusion with DNA microarray co-expression clusters in chapter 3, the name EST contigs is used throughout the entire thesis.

We previously showed that genes recognised as 1:1 orthologues in the genomes of *A. gambiae* and *D. melanogaster*, code on average for proteins with 56% sequence identity (Zdobnov et al., 2002). This suggested that many well-characterized proteins of *Drosophila* and other evolutionary more distant organisms such as human, share only limited identity with *Anopheles* proteins. This limited the utility of more comprehensive but automatically derived non-redundant protein collections such as UniRef90 and even UniRef50 (representing sequences merged at 90% and 50% sequence identity respectively), where best hits were dominated by poorly annotated predictions from genome sequencing projects. To capture such weak homologies we used the sensitive Smith-Waterman algorithm (Smith and Waterman, 1981) (as implemented by Paracel) to compare all forward translations of the EST contig sequences with sequences of known proteins from the manually curated UniProt/Swiss-Prot database. We then extracted from that database a concise annotation for the best matching sequence, identified with a E-value cut-off of less than 0.001. When available, we tentatively assigned Gene Ontology (GO) functional annotation terms (Ashburner et al., 2000) to the EST contigs, inferred from the best matching protein in UniProt/Swiss-Prot database. The UniProt/Swiss-Prot to GO mapping is provided by the GOA project at EMBL-EBI (Camon et al., 2004). The GO hierarchy was traversed in a 'bottom to top' manner to assign the high level 'GO-slim' functional classes, which can be further compared to the patterns of correlated expression as identified in the DNA microarray experiments. We also analysed the EST contig sequences for characteristic signatures of known protein domains, using state-of-the-art HMM profiles, as defined in PFAM and SMART (Bateman et al., 2004; Letunic et al., 2004) and summarised in InterPro.

We identified groups of orthologous genes between the predicted full proteomes of *A. gambiae* and *D. melanogaster* and broader orthologous groups including other animal genomes with full genome coverage using an Inparanoid-like (Remm et al., 2001) procedure. Orthologous genes were then used as markers to identify the conservation of the genomic arrangement (synteny) as described before (Zdobnov et al., 2002) using SyntQL tool (Zdobnov, unpublished).

Implementation

AnoEST was implemented as a relational database using mySQL software (http://www.mysql.com/). An interactive web interface to the data is provided using PHP (http://www.php.net/). EST clustering was implemented in Perl using a DBI interface to the mySQL backend, to allow scaling to higher numbers of sequences without additional computer memory requirements.

Microarray assessment of EST contig expression

We experimentally assessed the expression of AnoEST-derived contigs utilising the developmental dataset that is presented in detail in chapter 3 of this thesis. Briefly, the experimental design interrogated 8 different time periods of the entire *A. gambiae* life cycle, from embryos to adults. Hybridisations were performed against an artificially constructed standard reference (SR), containing all spots of the array. Four replicates (three biological and one technical – dye swap) for each time period were performed. Manual inspection and statistical measurements were used to assess spot quality based on signal intensity versus local background levels and spot diameter. Negative spiked-in controls were used to calculate global background levels, and only data above three standard deviations of background intensity levels were considered for further analysis. Data were loaded into GeneSpring v7.0 (Agilent Technologies, Palo Alto, CA, USA), and normalised with the intensity dependent (lowess) normalisation algorithm. After replicate averaging, we selected ESTs that had reliable measurements in at least 33 out of the 37 hybridizations and exhibited t-test P-value less than 0.05 in at least 2 of the 9 time points. These criteria led us to consider the expression of 15,135 ESTs as confirmed during mosquito development. To quantify the level of expression of these ESTs we considered the maximum intensity signal from the periods that were analysed.

Results and Discussion

A. gambiae *EST classification*

We collected from public sequence databases (Benson et al., 2004; Kulikova et al., 2004; Miyazaki et al., 2004) 215,634 *A. gambiae* expressed sequences (178,618 from 5′ sequencing and 37,015 from 3′ sequencing) originating from 179,955 clones. Of these sequences 211,468 were aligned to 593,349 regions on the nuclear or mitochondrial genome. For 203,812 expressed sequences, a unique genomic origin could be recognised. We clustered ESTs (assigned them into groups representing distinct expressed loci) using the genomic sequence as template, as described in the Methods section. This allows for a more specific assignment of ESTs into contigs, as it prevents merging of distinct gene loci due to chimeric ESTs or domains with highly similar sequences. Three types of contigs were distinguished: 1) T-contigs (Transcribed contigs) that had at least one supporting EST from this genomic locus, 2) N-contigs (with No uniquely matched ESTs) that shared regions of high sequence identity to EST/cDNA sequences but could not be confidently identified as expressed (this fraction also included recent duplications when the corresponding EST/cDNAs could have been derived from any one of the duplicated regions), and 3) U-contigs (Unaligned) of ESTs that failed to align to the genome. The derived EST contigs were further identified with current Ensembl (Birney et al., 2004) gene predictions, annotated with orthology/homology to known proteins and protein domains using sequence analysis techniques and tentatively associated with GO (Gene Ontology) and GO-slim functional categories (see Methods section).

The descriptive statistics of the AnoEST data are provided in Table 2.1, which includes the numbers of different types of contigs and their annotation with respect to the Ensembl, UniProt/Swiss-Prot (Apweiler et al., 2004) and InterPro (Mulder et al., 2003) databases. Of the derived T-contigs, 13,173 (61%) were supported by more than one EST each; 9,944 (75%) of these had a statistically significant hit to known proteins in Swiss-Prot. Although a single EST is commonly considered unreliable as evidence of expression (Okazaki et al., 2002), between one fourth to one half of the 8,305 EST singletons were supported by various sequence features indicative of protein coding genes: they accommodated a correct gene model according to

Ensembl gene predictions, encoded known protein domains, or had significant homologues in Swiss-Prot. As discussed below, we also used transcriptomic data to verify expression of a substantial fraction of T-contigs during mosquito development.

	Contigs *(with ≥2 ESTs)*	Ensembl	Swiss-Prot	InterPro
TCLAG	21,478 *(13,173)*	11,608 *(8,048)*	14,131 *(9,944)*	8,015 *(6,368)*
NCLAG	46,560 *(20,219)**	2,079 *(1,105)*	6,971 *(4,247)*	2,456 *(1,485)*
UCLAG	3,881 *(82)*	n.a.	1,236 *(52)*	199 *(25)*

Table 2.1. Descriptive statistics of *A. gambiae* EST contigs. Numbers refer to overlaps with the current Ensembl gene set (14,364 genes in total), homology to known proteins in the Swiss-Prot Knowledgebase (matching 9,791 sequences) and hits with protein domains in the InterPro database (matching 2,144 distinct domains). Numbers referring to contigs supported by at least two sequences are marked in bold. The numbers of distinct Ensembl genes overlapping with T-contigs and N-contigs are 9,639 (8,020), and 1,821 (1,076) respectively, as some Ensembl genes overlap with more than one EST contig. *35,660 *(17,143)* of N-contigs, i.e. 77% *(85%)* respectively, are contributed by only 863 ESTs, that are aligned with 50 to 191 distinct genomic loci.

In total, 11,608 T-contigs overlapped with 10,726 Ensembl gene models (out of 14,364 Ensembl predictions as of 10 Aug. 2004, v23.2b.1), indicating that, despite very strict clustering criteria, the analysis probably engendered a minor number of fragmentation artefacts. On average, the derived EST contigs overlapped with Ensembl gene models by about 920 nucleotides corresponding to 70% of the shorter loci; 2,695 contigs overlap Ensembl gene models by over 90%. Only 452 EST contigs had shorter than 20% overlaps; these probably derived from UTRs. Interestingly, 9,870 T-contigs (4,789 of which were supported by two or more EST/cDNAs) had no associated Ensembl gene predictions.

N-type contigs were quite different: they were twice as numerous but have only 1/6 as many Ensembl overlaps as did the T-type contigs (Table 2.1). Moreover, 35,660 (77%) of the N-type contigs were formed by only 863 ESTs, each of which was aligned to at least 50 distinct genomic loci. These likely represent transposable elements in *A. gambiae*, as 24,984 N-type contigs showed significant homology to known transposable elements in RepBase9.12 (Jurka 2000, http://www.girinst.org). In contrast, only 1,312 T-contigs (61 with confirmed expression, see below) were homologous to repetitive elements. 2,079 N-type contigs were currently annotated as genes. However, only 860 N-contigs corresponded to recently duplicated genes, 220 of which had a corresponding gene model. The portion of ESTs that failed to align to the nuclear or mitochondrial genome of *A. gambiae* (U-type contigs) constituted less than 3% of all sequences (Table 2.1). Some U-type contigs may correspond to as yet

unsequenced regions of the genome, while most of them are likely to be of erroneous origin (data not shown).

Singletons			Clusters with ≥2 ESTS	
Confirmed	No data	Fraction	Confirmed	No data
28.4%	40.9%	With Ensembl gene prediction	67.9%	59.4%
43.7%	51.8%	With homology in SWISS-PROT	79.2%	71.8%
20.0%	33.5%	With Ensembl and SWISS-PROT hit	63.8%	55.1%
71.6%	59.1%	No Ensembl	32.1%	40.6%
48.0%	40.8%	No Ensembl, no SWISS-PROT hit	16.6%	23.9%
100% (1379)	100% (6929)	Total	100% (6582)	100% (6591)

Table 2.2. T-contigs with and without supporting microarray expression data.

Analysis of T-contigs

We used both bioinformatic and transcriptomic methods to analyse in detail the category of T-type contigs, which represent the most prominent fraction of the mosquito genes. This dual analysis is summarised in Figure 2.1 and Table 2.2, separately for singletons (left) and for contigs with ≥2 ESTs (right). We categorised contigs as having or lacking corresponding Ensembl gene predictions, homologues in Swiss-Prot, or overlaps of Ensembl and Swiss-Prot hits. Transcriptomic analysis utilised a developmental dataset encompassing expression profiles in embryos, larvae, pupae and adult mosquitoes to highlight the fraction of contigs with experimentally verified expression and is presented in detail in chapter 3. These data were collected using 20,000-element cDNA microarrays (MMC1). Out of the 8,922 corresponding T-contigs we scored 1,379 singletons (17% of the total number of singletons) and 6,582 contigs with ≥2 ESTs (50% of such contigs), summing up to 90% of the microarray elements, as being significantly expressed in at least two developmental stages (see Methods).

First, we explored the question of whether contigs with verified expression but without annotation represented low-level transcriptional leakage or whether they were expressed at levels comparable to those of recognised genes. For this purpose we compared the distribution of log2-transformed values of expression, for T-contigs with and without Ensembl gene prediction and for the fraction of contigs with and without Swiss-Prot homologues. As shown in Figure 2.2, in both cases genes with and without annotation showed rather similar distributions, with only a small shift

towards lower expression values in the absence of annotation, which was slightly more pronounced for contigs with Swiss-Prot homologues. Only 61 T-type contigs with confirmed expression, 20 of which had a corresponding gene model, showed significant homology to *A. gambiae* transposable elements. This comparison suggested that most of the 3,100 EST contigs that are currently lacking a predicted gene model had detectable expression and were likely to be actual genes.

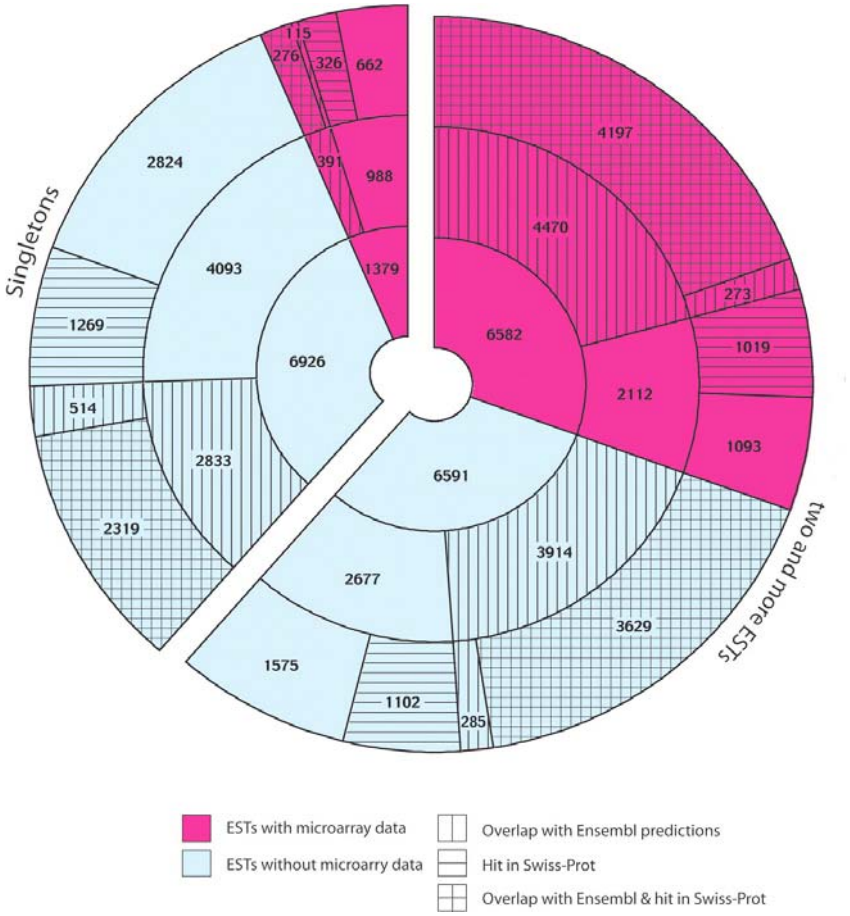


Figure 2.1.Analysis of the 21,478 T-contigs. The chart lists numbers of T-contigs of which expression during mosquito development was confirmed by microarray experiments (pink) and numbers of contigs for which microarray-based expression was not tested or detected (blue). Numbers are provided separately for contigs with two or more ESTs (right) and singletons (contigs with one EST, left). For each category the numbers of contigs with and without Ensembl gene predictions, as well as the numbers with and without homologous in UniProt/Swiss-Prot are indicated. The inner ring lists the total number of EST contigs with and without microarray data, and the outer two rings partition these contigs according to the associated annotation.

We then compared the T-contig subsets with verified expression with those lacking microarray data (mostly not represented on the microarrays). These subsets were reasonably similar in terms of presence or absence of corresponding Ensembl predictions, Swiss-Prot homologues, or both (Table 2.2). As expected the microarray-

expressed subset was substantially (5-fold) smaller than the subset lacking microarray data in the case of singletons, whereas the subsets were of equal size for ≥2 ESTs contigs. Surprisingly, the prevalence of Ensembl and Swiss-Prot hits was actually higher among singletons lacking supporting microarray data (Table 2.2).

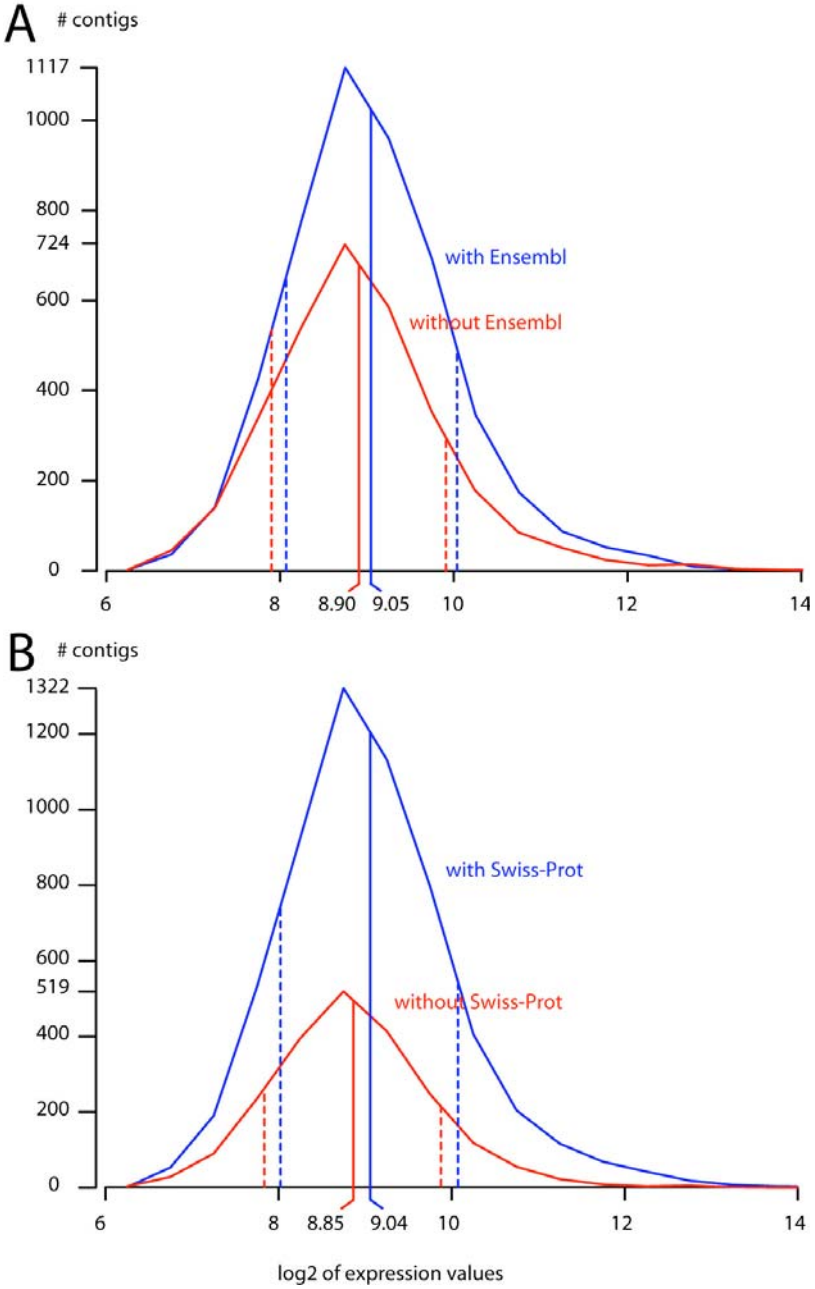


Figure 2.2. Comparison of distributions of T contigs with and without Ensembl or SWISS-PROT overlaps. A) Comparison of $\log_2$ expression value distributions for T-contigs with and without overlaps with Ensembl gene predictions. The graph also depicts mean (solid line) and standard deviation (dashed lines) values for the corresponding distributions. B) Comparison of $\log_2$ expression value distributions for T-contigs with and without homology hits in the Swiss-Prot knowledgebase; mean (solid line) and standard deviation (dashed lines) values are also shown.

Based on the analysis summarised in Figures 2.1 and 2.2 and in Table 2.2, our working hypothesis is that a substantial fraction of EST singletons represented actual genes, as do most of the ≥2 ESTs contigs. These data suggest that the number of genes in the *A. gambiae* genome may be substantially higher than currently predicted. A similar conclusion was recently drawn from another independent study of full-length cDNAs for *A. gambiae* (Gomez et al., 2005) and for the *Drosophila melanogaster* genome by using a combined bioinformatics and expression profiling approach (Hild et al., 2003).

Interface to the AnoEST database

The data discussed above were structured into a relational database, for which we developed a user-friendly Web interface, available at http://web.bioinformatics.ic.ac.uk:8080/. It allows querying for the EST/cDNA accession number, clone identifier, derived EST contig identifier, Ensembl gene identifier, Swiss-Prot accession numbers of homologous proteins and associated GO terms, permitting logical combinations and flexible regular expressions.

Examples of the available interactive searches are represented in Fig. 2.3. By default, the information on queried sequences is returned in a condensed format showing data corresponding to the best matching EST contig (Fig. 2.3A). The 'Sequences' tab at the top of the interface allows retrieval of the sequences in FASTA format and, if required, generates reverse complemented sequences, e.g. for 3′-sequenced clones. The 'Details' tab makes available more extensive information on similarity to known proteins and protein domains, orthology, GO and 'GO-slim' categories (Figure 2.3B). The annotation available for each corresponding genomic region in Ensembl can also be explored through a direct link to the genome browser. The 'Homology' tab refers to the full records of a similarity search of the EST contig consensus sequence against the UniProt/Swiss-Prot protein database. The records allow manual inspection of the alignments and provide html references to the corresponding entries in the UniProt/Swiss-Prot database.

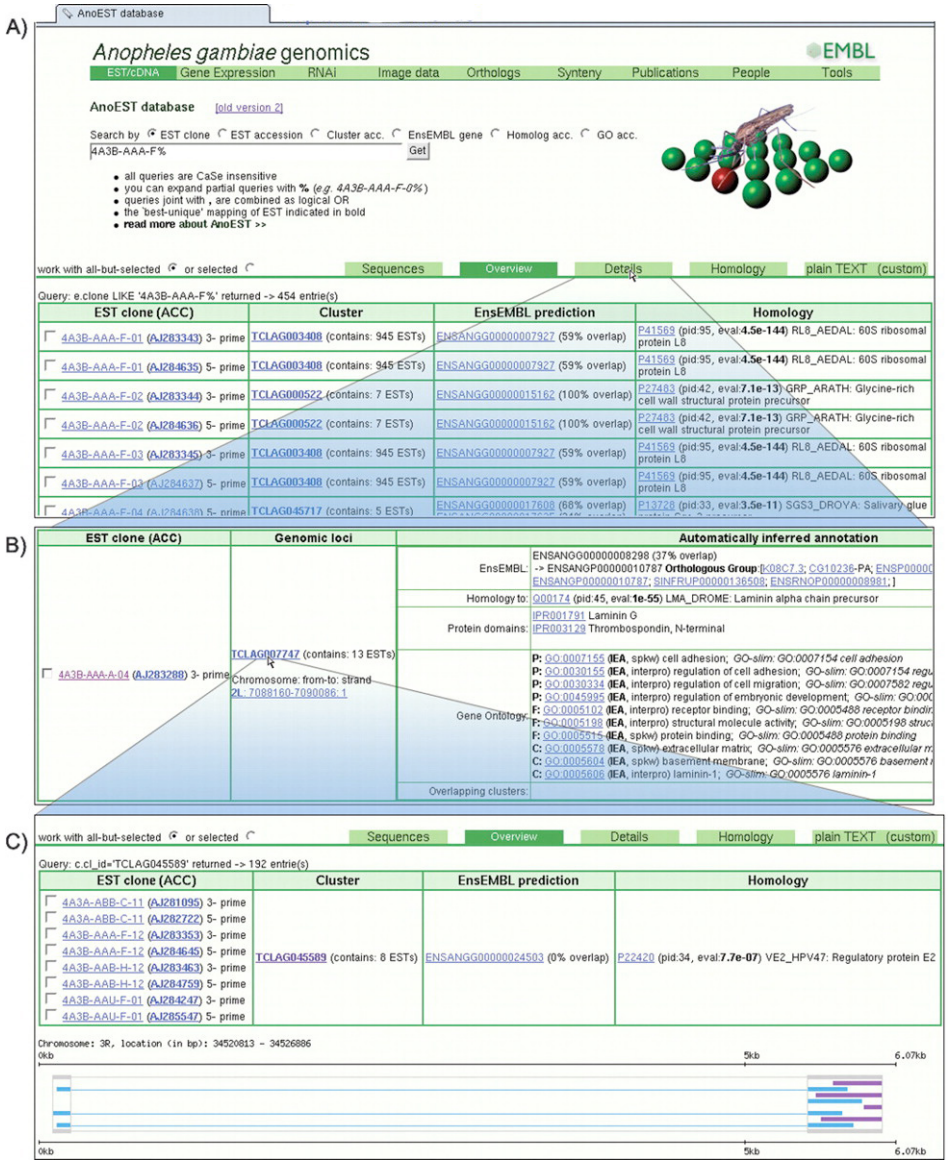


Figure 2.3. Interactive searches available in AnoEST. Searching with an EST clone identifier allows A) an overview of associated EST sequences, corresponding clusters, overlapping Ensembl genes, and best hit in the Swiss-Prot database and its description. B) The detailed view gives in addition coordinates of EST cluster match to the genome; links to orthologous groups identified on the basis of corresponding Ensembl gene predictions; protein domains collected in the InterPro database; and corresponding GO terms. When examining EST clusters, C) a graphical representation of overlapping ESTs permits visualization of the underlying exon-intron structure.

When exploring all expressed sequences assigned to one contig (Figure 2.3C), the visualisation of EST alignments to genome allows a quick grasp of the gene organisation, EST coverage and quality of the clustering. Sequences derived from 5′- and 3′-ends are coloured differently. The scale bar provided indicates the real contig length over the genomic alignment, sized to fit to the image. The EST contig image is mapped by html links to EST records for exploring cases of interest.

AnoEST utility for microarray analysis

AnoEST facilitated functional analysis of transcriptional data derived from existing cDNA microarrays by annotating microarray elements using information from the database via the EST identifiers. To allow further exploration of the annotation using specialised software or Excel spreadsheets, the microarray grid annotation (currently the 4K, MMC1/20K, and the full genome MMC2 microarray) is provided as tab-delimited text files. The constrains imposed by such representation limited the complexity of included data, e.g. each element is associated with only best matching EST/cDNA contig and its functional annotation. In order to draw solid conclusions for the expression of a genomic locus it is very important in DNA microarray experiments to evaluate all possible cross-hybridisations. That is possible utilising the 'CrossMapping' column included in the files that lists all contigs sharing high sequence identity for each of the microarray elements. Moreover, as the microarray elements enclose both DNA strands that could potentially contribute to the spot signal, we reported all EST contigs that are on opposite strands but overlapped by at least 60nt. This is summarised in the 'Overlapping contigs' column, generated by concatenating the corresponding contig identifiers ordered by significance of sequence homology to known proteins, which allows users to easily recognise and group them.

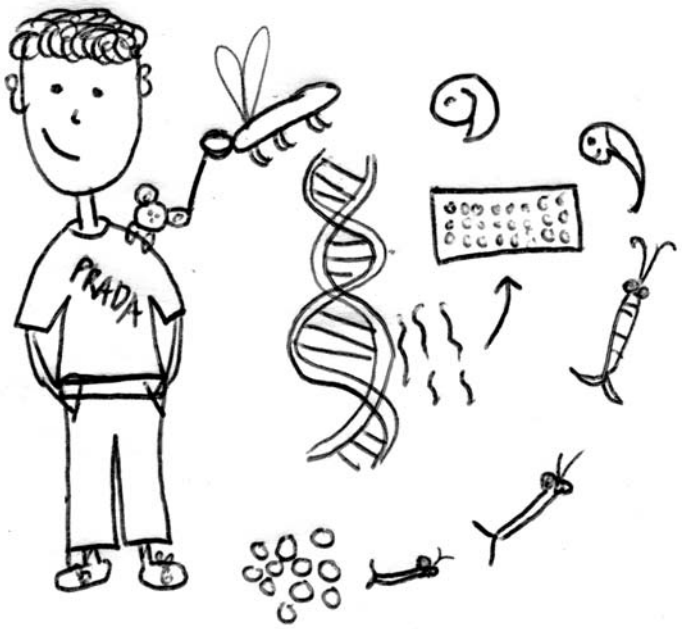
Conclusions and future development

The publication of the genome sequence of *A. gambiae* has led to the analysis and rapid organisation of available genomic information to a number of mosquito public databases. Ensembl currently holds genomic information, AnoBase holds biological information about *Anopheles spp.* and other databases hold diverse information, for example AnoXcel for proteomic data (Ribeiro et al., 2004). AnoEST, the newly developed database that is reported in this thesis is a database of expressed sequences of *A. gambiae.*

To construct this database, we mapped the expressed sequences to the genome, using the sequence of *A. gambiae* as a template, and constructed EST contigs from these overlapping expressed sequences. Those EST contigs have been supplemented with inferred functional annotation including similarities to known proteins, protein domains, and Gene Ontology (GO) functional categories. Using the MMC1

microarray data in conjunction with AnoEST, we have experimentally confirmed expression of 7,961 contigs during mosquito development. Of these, 3,100 are not associated with currently predicted genes. Moreover, we found that contigs with confirmed expression are non-biased with respect to the current gene annotation or homology to known proteins, and consequently we expect that many of the unconfirmed contigs are likely to be actual *A. gambiae* genes. However, besides the identification of previously undiscovered genes, AnoEST is a vital resource for the interpretation of expression profiles, which are derived using DNA microarrays.

AnoEST was initially developed as an independent database but it has already been adapted to serve as one of the functional genomics modules of a new integrated genomic data resource, VectorBase (http://www.vectorbase.org). VectorBase is a collaborative effort of five institutions to bring together a series bioinformatics databases with the aim of developing an integrated resource for insect disease vectors, which include *A. gambiae*, *Anopheles funestus*, *Aedes aegypti* (yellow fever mosquito), *Culex pipiens*, *Glossina spp.* (tse tse flies), *Rhodnius prolixus* (triatomine bugs) and the tick *Ixodes scapularis.* At the time of writing of the thesis, the draft sequence of the *Aedes aegypti* genome had been released to the public and a new database for expressed sequences, AedEST, had been created, utilising the same algorithms and protocols developed for AnoEST. AnoEST will be regularly updated to reflect changes in the gene annotation and include new sources of expressed sequences. Future developments include further refinement of the EST to genome mapping, full automation of the methods, as well as integration to the gene prediction algorithm of Ensembl. In addition, a more thorough analysis of the TCLAG contigs that do not overlap with Ensembl being models is being sought.



Chapter 3

Transcriptomic analysis of the life cycle of the mosquito *Anopheles gambiae* and its comparison to the *Drosophila melanogaster* life cycle

Introduction

The A. gambiae lifecycle

Mosquitoes have been the focus of entomological research for many decades and numerous morphological, systematic and disease-related studies have been conducted. They were mostly studied in the context of the many natural history studies that characterised the zoological interest during the period between the second half of the eighteenth and first half of the ninteenth century. As a result, early studies in the life cycle of the mosquito species have focused in the understanding of the basic biology and factors that influence mosquito growth and behaviour. Those early works are reviewed in many excellent textbooks (Bates, 1949; Christophers, 1960; Clements and Clements, 1992) and have provided the framework for the current understanding of mosquito biology.

In *A. gambiae,* the blood feeding of an adult female that initiates egg development can be considered the start of the lifecycle. Digestion of blood proteins yields aminoacids, which are reconstituted in the mosquitoes' fat body, transported into the ovaries and incorporated into the oocytes. Eggs do not mature continuously but in batches, following the periodic blood meals; approximately 100 or more eggs mature from each blood meal. The role of the males to inseminate the females, which is done either when a female mosquito enters a swarm of males or by individual mating outside swarms. The females are believed to mate only once, as they can store sperm in their spermathecae to fertilise many batches of eggs. The final act of mating is the injection of a substance from the male, which blocks the passage of sperm from any subsequent copulations, although there are indications that this phenomenon does not persist for the entire life of the female mosquito.

After mating and blood feeding, approximately 50-500 eggs are fertilised while traversing the genital chamber and are deposited on water or sites that will be flooded. Embryogenesis starts almost immediately after oviposition and normally lasts two to three days. Like many other insects, mosquitoes exhibit 'complete metamorphosis'; from the embryo to larva, to pupa and then to adult. The larva hatches once it is being formed and can survive for a few days in the absence of water. Four larval instars ensue accompanied by continuous growth. They live in an aquatic environment and feed by filtering bacteria, diatoms, algae and other particles

from water. In that period, larval organs are functional and adult organs incipient or slowly developing. Most of the organs are histolysed during metamorphosis (from the larva via the pupa to the adult), while others persist. When the adult body has fully formed, a pressure difference along the pupal cuticle results in the rupture of the structure and the release of the organism. The adult rests at the water surface for some time, as it will need to dry out and harden before it can fly. Subsequently, mating will be the first activity of the young adult *Anopheles.*

Although much information is available about the timing of the *A. gambiae* lifecycle, detailed molecular and cell morphological studies are lacking. For this reason, studies in the molecular level in *A. gambiae* – as in other insects – have largely used *Drosophila melanogaster* as the model system, the lifecycle of which has been analysed in considerable molecular level and information is readily available. *Anopheles* and *Drosophila*, which are believed to have diverged from a common ancestor ca. 250 mya, have comparable lifecycles. However, *Anopheles* has four larval stages compared to three in *D. melanogaster* and some notable differences at both the morphological (Monnerat et al., 2002) and molecular (Goltsev et al., 2004) levels have been observed. Thus, a large-scale molecular analysis of the lifecycle of *A. gambiae* could potentially analyse similarities and differences between those insects and give information on their adaptation to different environments.

DNA microarrays as tools in basic biology research

DNA microarrays are important tools for large-scale gene expression profiling studies. The early 'precursors' of the modern microarrays have been bacterial cosmid libraries spotted on nylon membranes, which were used for hybridisation analyses. A series of advances both in chemistry, with the manufacturing of robust fluorescent dyes, and in engineering, with the construction of robotic devices of great precision, has facilitated the miniaturisation of such arrays and the development of glass slide printed microarrays.

The first such microarray reported in the literature was a cDNA microarray of 1,161 DNA elements of human genes, which was used to analyse gene expression in a cancerous cell line (DeRisi et al., 1996). Since then, advances in microarray manufacturing, experimental protocols and analysis platforms (reviewed in (Stoughton, 2005)) have resulted in the construction of DNA microarrays for most

model organisms. Examples of applications in these organisms include gene expression profiling during sporulation and the cell cycle in the budding yeast *Saccharomyces cerevisiae* (Cho et al., 1998; Chu et al., 1998), the lifecycle of the nematode worm *Caenorhabditis elegans* (Hill et al., 2000) and the fruitfly *D. melanogaster* (Arbeitman et al., 2002), the mitotic cycle of human fibroblast cells (Cho et al., 2001) and seed development in the plant *Arabidopsis thaliana* (Girke et al., 2000). The microarrays have thus proven to be valuable tools for the study of the expression profiles of numerous model organisms and their use could potentially be extended to less studied ones.

DNA microarrays in A. gambiae

Our laboratory has been paramount in the development of DNA microarrays in *A.* gambiae. The first microarray that was constructed contained 3,840 ESTs representing approximately 2,500 genes (Dimopoulos et al., 2000). This platform was used to explore mosquito immune responses against bacteria and *Plasmodium* parasites (Christophides et al., 2002; Dimopoulos et al., 2002) and to identify differences of gene expression between *Plasmodium* refractory and susceptible mosquitoes (Kumar et al., 2003). Other available microarray platforms for *A. gambiae* include: a genome-wide oligonucleotide platform by Affymetrix and a custom cDNA microarray to assay gene expression in blood feeding (Dana et al., 2005; Marinotti et al., 2005); a small scale 'detoxification' microarray for the study of metabolic basis of insecticide resistance (David et al., 2005); a custom microarray for the identification of odorant binding proteins (Biessmann et al., 2002) and two new platforms that have been developed from our laboratory, the MMC1 cDNA microarray, which contains 19,680 ESTs that correspond to approximately 8,872 EST contigs, and the whole genome MMC2 microarray platform.

Aims of the present study

In depth molecular studies of the lifecycle of *A.* gambiae are of significant importance in understanding mosquito biology. The present study reports the construction of the MMC1 microarray platform and its use to describe genome-wide expression profiles during the *A. gambiae* lifecycle. The aim of this study is to reveal

transcriptional programmes that are associated with critical developmental transition stages and define similarities in major gene functional classes in the mosquito lifecycle. It also focuses on the determination of temporal gene expression in four different tissues of the adult female.

Furthermore, a previous genomic analysis identified a large percentage of gene orthology between *Drosophila* and *Anopheles* (Zdobnov et al., 2002). A developmental transcriptomic study in *Drosophila* allowed us to extend the analysis to include information about orthologous gene expression. This study reports the first large scale comparative transcriptomic analysis between diptera in which we demonstrate a strong correlation between expression patterns of orthologous gene pairs and explore similarities and differences in gene expression profiles that underlie the variant life styles of *Anopheles* and *Drosophila*.

Materials and methods

EST library construction, sequencing and clustering

Two oligo-dT primed cDNA pools were constructed from pooled developmental stages (NAP1 library) and adult heads (NAH library) of a *Plasmodium* susceptible 4a r/r strain of *A. gambiae*. Sequences were directionally cloned in the pT7T3D-Pac plasmid vector (GE Healthcare, UK) with modified cloning sites. Plasmids were isolated from the NAP1 and NAH libraries, gel-purified to remove empty vectors and the resulting libraries were normalised as described elsewhere (Bonaldo et al., 1996). After transformation of 15,513 plasmids into *E. coli* DH10B cells, single bacterial colonies were sequenced via the T7 promoter sequence (cDNAs 5'-end) using the MEGA Base system (GE Healthcare, UK). The sequences were cleared of vector contamination and submitted to the EMBL DNA sequence database.

For *in silico* clustering analysis, we have used the annotation information of AnoEST, which has been described in detail in chapter 2. Briefly, the MMC1 EST sequences together with all 200,121 additional *A. gambiae* cDNA and EST sequences deposited in public databases, were retrieved and channelled into the AnoEST clustering pipeline (Kriventseva et al., 2005), mapped to the *A. gambiae* genome sequence and assembled into EST contigs, These contigs were supplemented with functional annotation information which included overlapping Ensembl gene models, orthologous genes in other species (including *D. melanogaster*), homology to known proteins, INTERPRO protein domains, associated Gene Ontology terms and corresponding classification into broad GO-slim functional groups. Additional annotation about gene families was retrieved from Ensembl database.

Microarray construction

Microarrays were constructed with 15,840 ESTs from the NAP1 and NAH libraries and 3,840 ESTs from the 4a3A and 4a3B library, which were previously used for the construction of the first mosquito DNA microarray (Dimopoulos et al., 2000; Dimopoulos et al., 2002). EST sequences were amplified from bacterial cultures using T3 and T7 primer sequences, derived amplicons were cleaned with the Nucleospin® Robot-96 Extract kit (Macherey-Nagel, DE), concentrated to 150-300

ng/μL, resuspended in 3x SSC buffer and spotted onto glass slides coated with N-16-aminohexyl aminosilane. This new platform was annotated as MMC1 (Mosquito Microarray Consortium 1) and the coverage of the platform on the *A. gambiae* genome is shown in Fig. 3.1. Based on the AnoEST analysis (v. 5), the 19,680 MMC ESTs are encompassed in 8,872 EST contigs, each with unique hits on the *A. gambiae* genome (TCLAG contigs); of these, 5,367 (ca. 60%) overlap with Ensembl-built gene models. Our analysis excluded UCLAG and NCLAG contigs that are either unmapped or repetitive and map at multiple genome loci.

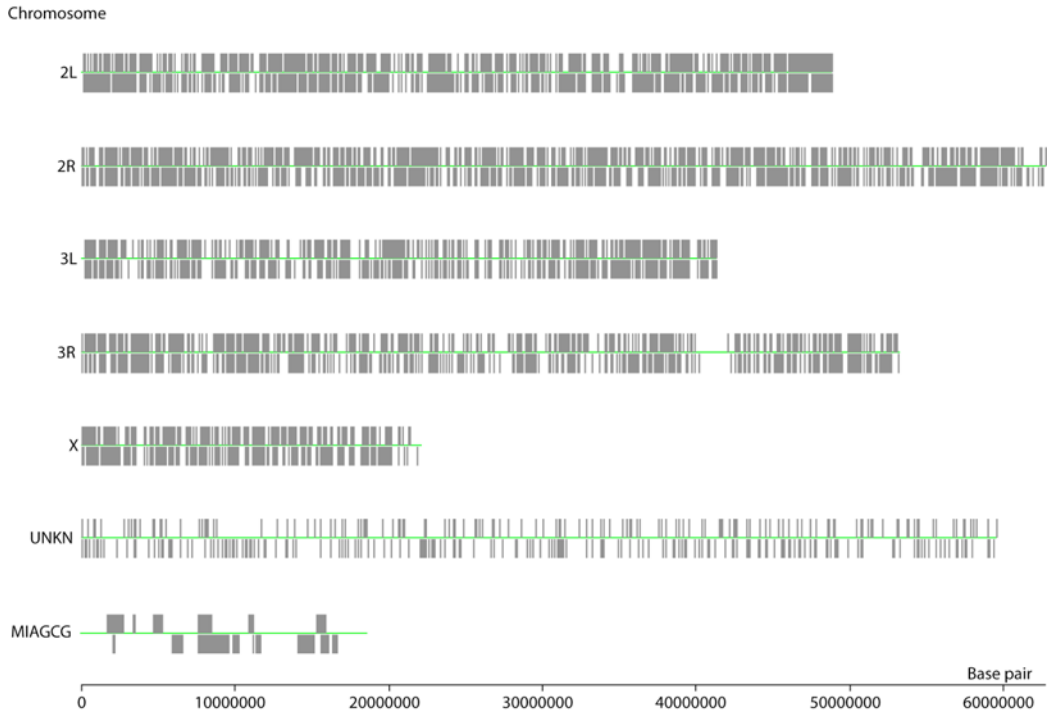


Fig. 3.1. Coverage of the MMC1 microarray to the A. gambiae genome. Picture shows the MMC1 ESTs mapped to the *A. gambiae* chromosome arms, the unknown chromosome (UNKN, representing contigs that have not been aligned could not be positioned in the genome assembly) and the mitochondrial (MIAGCG) genome sequence. ESTs were mapped either in the sense (above) or antisense (below) orientation. The mitochondrial sequence has been magnified 1200 times.

Mosquito rearing and preparation of experimental RNA samples

A. gambiae mosquitoes of the G3 strain were reared at 27° C at a 75% relative humidity with a 12/12-hour cycle. Adult mosquitoes were fed on wet cotton pads supplemented with a 15% sucrose solution. Larvae and pupae were raised in 0.1% salt water and fed with powdered cat food or cat food pellets. To initiate egg production, females were allowed to feed on sedated mice and 2-3 days after blood meal, eggs were collected from water-soaked Whattman papers and placed in 0.1% salt solution.

The procedure to generate each microarray sample is summarised in fig. 3.2. Total RNA was extracted using Trizol reagent (Invitrogen, Carlsbad, CA) from: embryos, 50-350 larvae, 30-40 adults and 30-42 adult tissues. Following quality inspection by gel electrophoresis and normalisation of all RNAs to 1 μg/μL, 5 μg samples were subjected to first and second strand cDNA synthesis (Invitrogen, CA, USA), using primers that incorporated the T7 promoter sequences (TAATACGACTCACTATAGGG) at the 3' end, after the poly(T) tail of all cDNAs. Amplification was performed with T7 transcription reactions, which were used to produce complementary mRNAs (cmRNAs).

Preparation of standard reference RNA

To generate standard reference (SR) *in vitro* as previously described (Sterrenburg et al., 2002), small amounts (1-10 ng) of all PCR products were synthesized, pooled and purified twice by Phenol-chloroform extractions, followed by ethanol precipitations and resuspension in 5mM TrisCl pH 8.0, 1 mM EDTA (Fig. 3.3) Those products were used as templates to produce complementary RNA (cmRNA) using multiple pools of *in vitro* transcription reactions (MEGAscript® T7 kit, Ambion, USA).

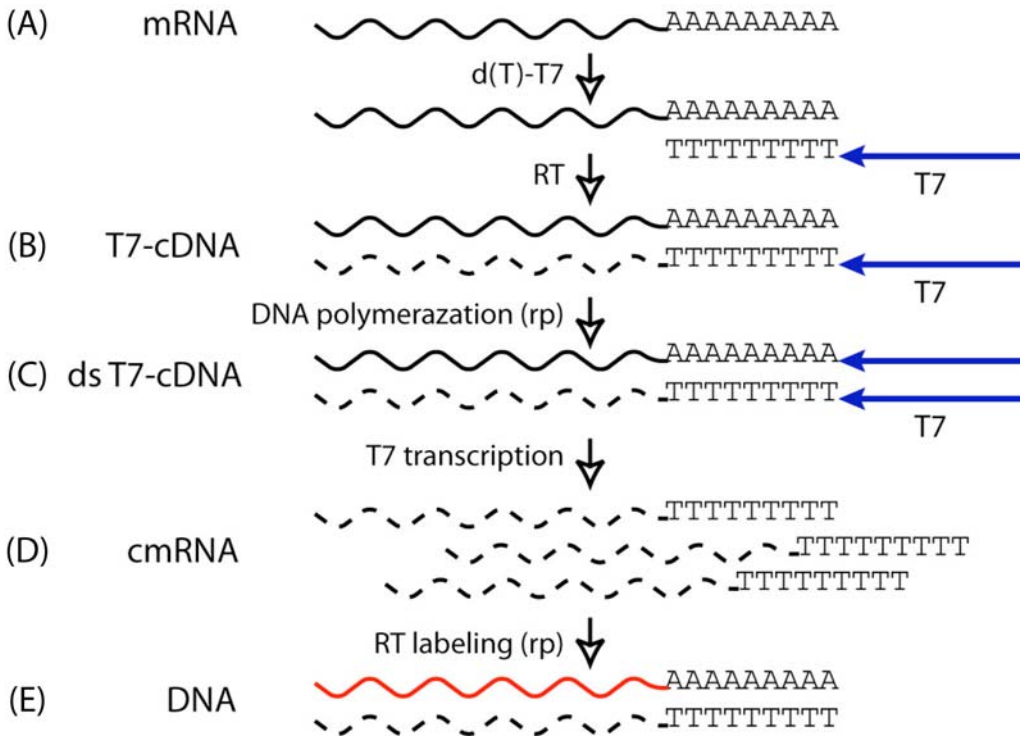


Fig. 3.2. Schematic representation of the amplification and labelling of the microarray experimental samples. Initially, the mRNA (A) is used as a template for the synthesis of a complementary strand of DNA (first strand synthesis, B) by annealing a poly (T)-T7 oligonucleotide. Subsequent RNAse treatment, DNA polymerisation and DNA ligation results in degradation of the mRNA strand and construction of a complementary DNA strand (second strand synthesis, C). This product is used for T7 transcription reactions to produce multiple copies (amplification step, D) of cmRNA, which is complementary to the initial mRNA strand population. cmRNA will be subsequently used to produce a labelled complementary strand (labelling, E) by using fluorescent-labelled dUTP. (Fig. Courtesy of G.K. Christophides).

Microarray hybridisations, image and data analysis

Hybridisations were performed as described (Christophides et al., 2002; Dimopoulos et al., 2002), using 5 μg cmRNA of samples 5 and 100 ng SR sample that were labelled with Cy-5 and Cy-3 dyes, respectively (GE Healthcare, UK). One dye-swap experiment was performed for each experimental set. After hybridisation and washes, microarray slides were scanned using a GenePix 4000b (Molecular Devices, Sunnyvale, CA, USA) semiconfocal scanner, and image analysis and measurements were performed with GenePix Pro 3.0 software (Molecular Devices, Sunnyvale, CA, USA). Initial spot evaluation was performed by visual inspection, and derived measurements were subjected to strict filtering. The criteria required that fluorescence intensity in each channel is above 1.5 times the background intensity (local background filtering) and above the average intensity plus three standard deviations of the negative spike-in controls (Lucidea Universal ScoreCard®, GE

Healthcare, UK, global background filtering) and that spot diameter is neither higher nor lower than three standard deviations of the average spot diameter. Filtered data were loaded in a custom-made microarray annotation platform into GeneSpring version 7.0 software (Agilent Technologies, Palo Alto, CA, USA) To perform dye swap, signal and reference channel measurements were reversed. For data normalisation, we used a local weighted linear regression (Lowess) algorithm. In brief, a Lowess curve was calculated each time utilising 20 percent of the data, fitted to the log-intensity vs. log-ratio plot and the resulting normalising curve used to adjust the standard reference value for each measurement. This is the standard normalisation process for two-colour microarray experiments that minimises effects due to differential dye properties and allows for comparisons among hybridisations in different microarray slides.

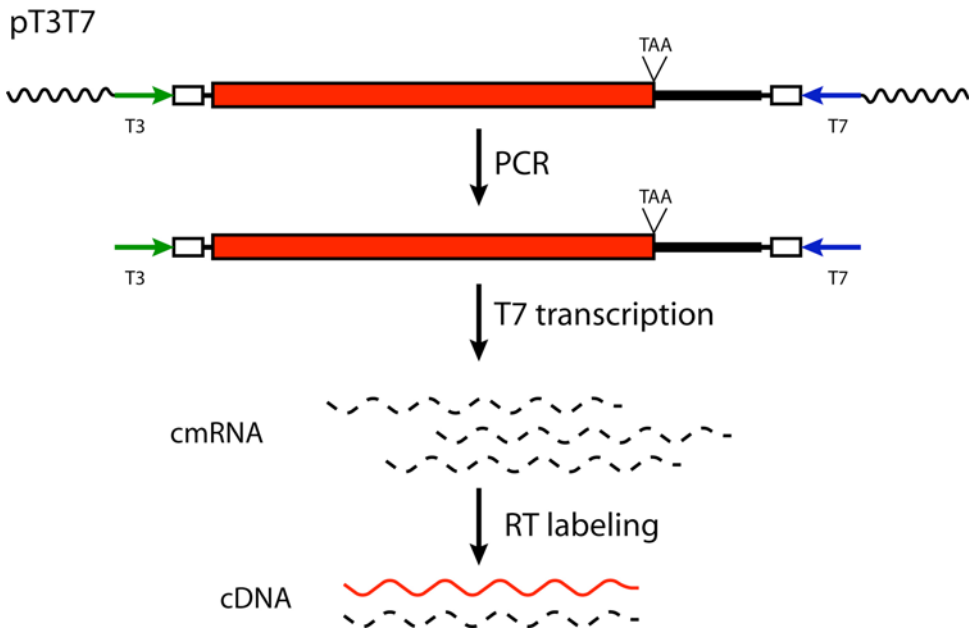


Fig. 3.3. Schematic representation of the construction and labelling of the SR sample. Fragments of the all the spots contained in the microarray were produced by PCR reactions from the initial pT3T7 plasmids. Subsequent T7 transcription reactions produced ssRNA (cmRNA), as described in the second stand synthesis procedure (Fig. 3.2). The cmRNA was then used to produce a labelled complementary strand by utilising fluorescent-labelled dUTPs. (Fig. courtesy of G. K. Christophides)

For producing co-expression clusters, the datasets have been exported from GeneSpring to specific gene clustering programs. Genecluster version 2.0 (http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html) was used for the SOM clustering (Tamayo et al., 1999; Toronen et al., 1999; Vesanto and Alhoniemi, 2000) of the developmental programmes and Cluster version 3.0 (http://bonsai.ims.u-

tokyo.ac.jp/~mdehoon/software/cluster/software.htm) for k-means clustering of the tissue patterns (Eisen et al., 1998). SOM clusters were visualised and processed in R-statistical package version 2.1.0 ((Team, 2006), http://www.r-project.org/) and k-means clusters were visualised with the Treeview Java version 1.0.13 programme (http://rana.lbl.gov/EisenSoftware.htm).

To detect global patterns of expression related to functional protein annotations, we grouped TCLAG contigs based on either Gene Ontology (GO) terms or INTERPRO domains. For each group and time point, we defined the percentage of TCLAG contigs showing top or bottom expression (upper 25% and lower 25% of the expression range, respectively); the low expression percentages were then subtracted from the respective high expression percentages. Only GO terms and INTERPRO domains containing at least 20 distinct TCLAG contigs were included in the analysis. The dataset of subtracted group-specific percentage values for each term or domain was subjected to k-means clustering.

Comparative transcriptomic analysis

The raw microarray data from a *D. melanogaster* lifecycle study (Arbeitman et al., 2002) were downloaded from http://genome.med.yale.edu/Lifecycle/Data_download/. Briefly, this study interrogated 73 different time periods in Drosophila: 31 overlapping embryonic time periods, 10 larval, 18 pupal and 14 for adult male and female. Spots were subjected to the same criteria as for the *A. gambiae* dataset, except for the criterion relating to negative spike-in controls, due to the absence of these controls in that microarray. The expression profiles of individual ESTs corresponding to the same *Drosophila* gene were averaged. Data were normalised under the Lowess curve algorithm, as described previously, and 3,571 genes with reliable measurements in at least 130 of the total 151 hybridisations and with a t-test P-value less than 0.05 in at least 1 of the developmental time periods of the *Drosophila* study were considered for further analysis.

For comparison, the *Anopheles* and *Drosophila* lifecycle datasets were divided into an equal number of notional developmental periods. Newly emerged (ca. 12h post emergence) *Anopheles* males (M) and females (F) were compared with 24h-old *Drosophila* adult males (Am24h) and females (Af24h), respectively. Averaging of the embryonic and pupal time samples was based on a correlation analysis (data not

shown). Three (early, middle, late) comparable phases of larval development were defined from the five larval time periods of the *Anopheles* study and the ten of the *Drosophila* study, by using a sliding window procedure.

Orthologous gene pairs between *Anopheles* and *Drosophila* were constructed from best reciprocal hits and information from syntenic regions (Zdobnov et al., 2002). From this list a combined expression matrix of 1,039 unambiguous orthologous pairs was normalised to the median of each gene and the 50^{th} percentile of each microarray. Correlation analysis with Pearson and smooth coefficients was performed in GeneSpring and graphics were plotted using the R statistical package (Team, 2006). Subsequent analyses of gene groups with the same GO terms, INTEPRO domains or gene expression clusters were performed in the R statistical package using the skewness distribution measurement from the moments package. The Wilcoxon rank sum test (U test) was used for testing differences in mean values of the distribution curves.

Results

Experimental design

An initial *A. gambiae* population (P0 generation) consisted of approximately 400 adult female and male mosquitoes of the laboratory *G3* strain The females were allowed to feed on mice at days 3 or 8 of adulthood and correspondingly are defined as the experimental generations P1 and P2. The P3 generation was the progeny of P1 mated females that were fed on mice three days post emergence. The mosquito developmental lifecycle was sampled empirically at eight successive time periods (Fig. 3.4). These included embryos, five larval samples, pupae and adult females and males. Female adult tissues from head, gut, and carcass (the latter corresponding to the remnant after removal of head, gut, wings and legs) were collected from 1-day old females, and ovaries from 5 to 6-day old females that were blood fed 48 h earlier.

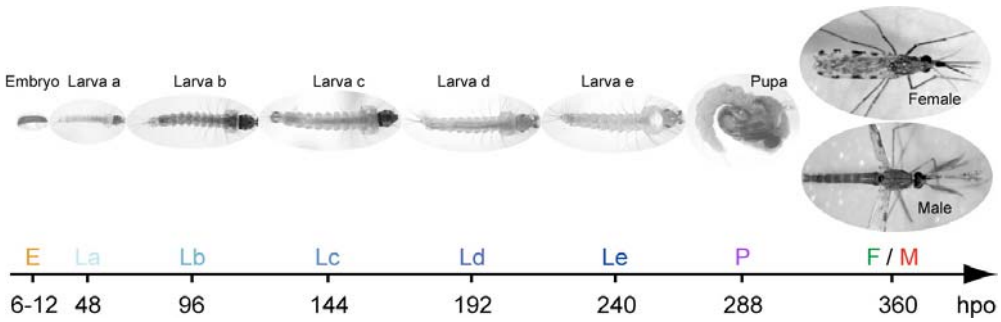


Fig. 3.4. Lifecycle microarray experiment. For the lifecycle expression analysis, samples were collected at eight time periods as indicated on the time scale in hours post-oviposition (hpo). These include embryos, E, collected 6-12h hpo; larvae collected at five periods: La approximately 48 hpo and the rest, Lb-Le, at 48 h intervals thereafter, irrespective of larval size and instar; pupae, P, collected 6-12 h after Le and freshly emerged (12-16 h) adult females, F, and males, M.

To investigate genome expression during mosquito development and in adult tissues, we performed competitive two-dye hybridisations of experimental and standard reference (SR) RNA samples to MMC1/20K EST microarrays. SR RNA was produced *in vitro* from all spotted ESTs. This mixture was utilised to provide consistent, non-zero reference values for almost all probes of the array, allowing us to effectively normalise all experiments. Dye-swap replicates were performed for all P1 samples. The expression data of ESTs mapping to the same EST contig were averaged. Only contigs that have single hits (TCLAG or T-contigs) on the currently assembled mosquito genome were processed further. Additional data will be usable

when the genome assembly is improved and identification of some additional EST clusters with newly recognised genes is achieved.

Pairwise comparisons of the SR samples in both the developmental timing and adult tissue studies demonstrated high reproducibility (Fig. 3.5). Next we examined the reproducibility of TCLAG contig expression (in three biological and one technical dye swap replicates) using a conditional tree analysis with Pearson correlation coefficient. The reproducibility was high in the developmental study as shown by the cohesive clustering of embryonic, pupal, female and male adult replicates and all larval replicates combined (Fig. 3.6). However, expression patterns at different larval periods were insufficiently distinct and there was some tendency for larval expression to cluster by mosquito population rather than time period; these features presumably reflect a predominant and inherent similarity of development in all larval instars and possibly some physiological differences in different mosquito populations. Inherent similarity of larval samples is consistent with the known continuous growth of both larval and imaginal disc tissues in the larva, and with the recurrence of developmental phases within each instar (e.g. a shift from growth to hormonally-induced epithelial retraction from the old cuticle and laying down of a new cuticle). We did not attempt to define separately such within-instar phases.

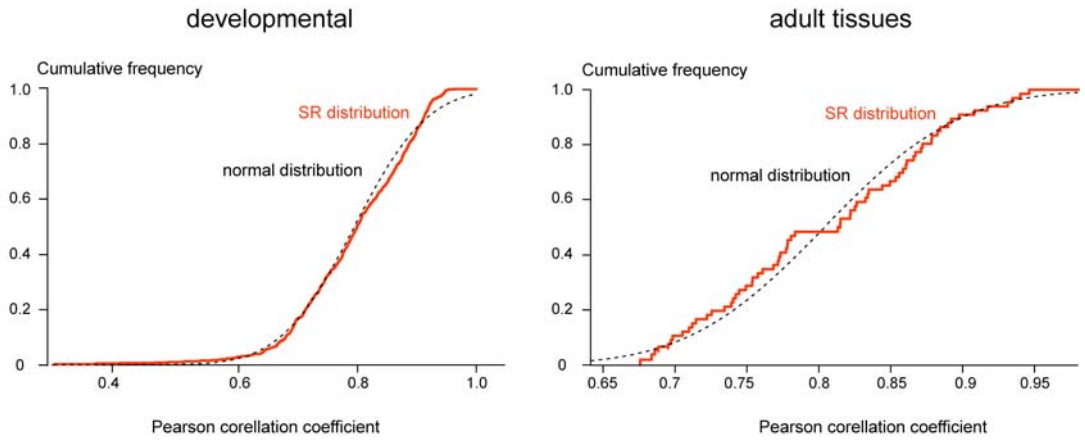


Figure 3.5. The SR sample comparisons. Cumulative distribution of the Pearson correlation coefficients for pairwise comparisons between SR and developmental (left) or the adult tissue (right) samples is shown in red. Dotted lines depict the normal cumulative distribution curves in both panels. Note that the vast majority of sample comparisons show correlation coefficients higher that 0.7.

Gene expression differences during development

To assess statistically significant changes of gene expression, we used a one-factor analysis of variance (ANOVA). Fig. 3.7 displays the number of TCLAG contigs that

show notable expression regulation at different P-values in both the developmental and tissue studies. The P-value of 0.001 admits numerous (2,421) differentially regulated TCLAG contigs (25% of the total) without many expected false positives (approximately 16 TCLAG contig in the case of P-value less than 0.001) and was thus chosen for analysis. Nearly two thirds of these (1,571) showed at least two-fold differences between their respective minimum and maximum expression values.

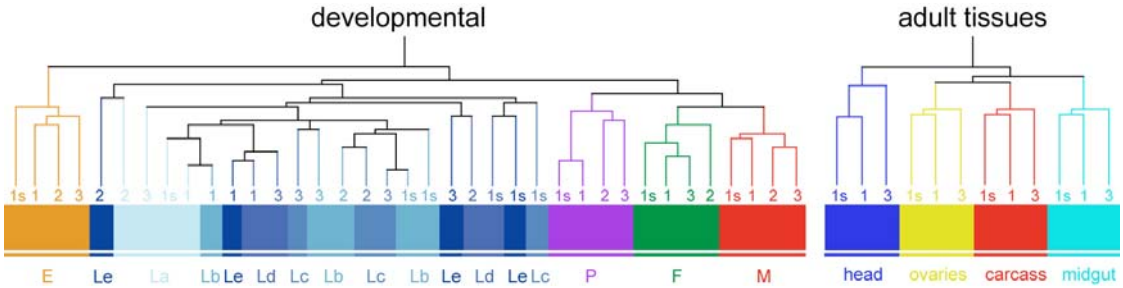


Fig. 3.6. Comparison of the hybridisation samples. Conditional tree of the developmental (left panel) or adult tissue samples (right panel) based on total gene expression patterns. Note the cohesive clustering of embryonic, pupal and adult female and male samples; larval samples other than La are intermixed.

We analysed the differential expression of these 2,421 contigs by pairwise comparisons of our 9 stages (Tukey test, Table 3.1 and Dataset[3] 3.D1). Consistent with the conditional tree analysis (Fig. 3.6, left panel), the embryonic expression profiles were by far the most distinct, displaying the highest number of differentially expressed contigs relative to other developmental stages. Pupal, male and female patterns were also quite distinct. In contrast, the larval phases were mostly indistinct and intermixed, although there was a tendency for late larval stages (Ld-e) to cluster together apart from earlier larval stages, (La-c). It is known that the precursors of adult organs (imaginal disks) develop continuously in larvae; their cell divisions are slow at early instars and acellerate at later instars. Of the few TCLAG contigs that showed clear differences between larval stages, several encode proteins implicated in nucleic acid binding, protein metabolism and cuticular constituents.

[3] The dataset contains the expression matrix as well as the annotation information of the corresponding contig lists and is available in the accompanying DVD-ROM.

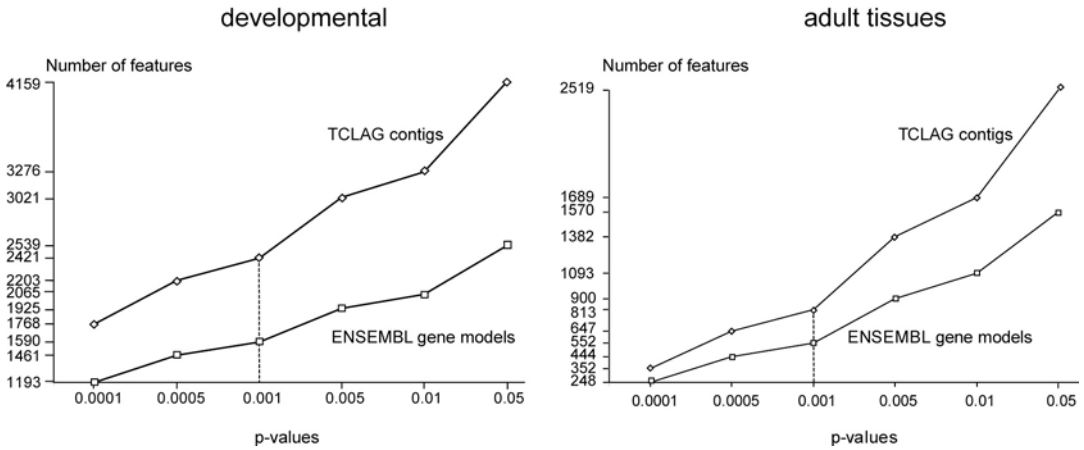


Figure 3.7. Differentially expressed contigs. Graph showing the number of TCLAG contigs and Ensembl genes that are differentially expressed at different ANOVA P-values in the developmental (left) or adult tissue (right) experiments. The cut-off $p \leq 0.001$ was used for subsequent downstream analyses.

A substantial number of TCLAG contigs differed in expression between pupae versus adult females (265) or males (245). Of these, many corresponded to components of the proteasome complex, reflecting the extensive histolysis of the larval tissues during pupal life; others encoded structural components of the cuticle, which is evidently different between pupae and adults.

We analysed the developmental transcription programmes using two complementary strategies. In the first, we performed automated clustering of the expression profiles of TCLAG contigs using self-organising maps (SOM) and then queried what functional categories are associated with each SOM expression cluster (Fig. 3.8). In the second strategy, we used a converse approach: we grouped the EST contigs by functional categories (GO annotation and INTERPRO domain content) and determined for each category the prevalence of high and low expression levels at each developmental stage; thus we identified differential RNA prevalence profiles, which were finally k-mean clustered (Fig. 3.9).

Developmental transcription programmes

We clustered the developmental expression profiles of the >2-fold regulated 1,571 ANOVA contigs, which included 1,065 Ensembl genes and 783 *Drosophila* gene orthologues, with a 5 x 6 SOM node that yielded distinctive and tight expression clusters (Fig. 3.8). Additional nodes did not produce any fundamentally new patterns. After careful consideration, these 30 clusters were combined (despite minor differences) into nine broad transcription programmes, some of which were

subdivided into sub programmes. The gene contents of these programmes are summarised below and in Supplementary Table 3.S1; the complete analysis of gene contents and their expression data is provided in dataset 3.D2.

La	Lb	Lc	Ld	Le	P	F	M
663(492)	678(513)	672(513)	816(626)	627(489)	818(626)	624(505)	1,009(759)
	14(12)	19(18)	68(59)	81(65)	279(232)	272(232)	506(384)
		0(0)	49(42)	61(51)	294(241)	281(238)	572(435)
			25(24)	37(31)	226(192)	214(191)	454(359)
				13(11)	333(296)	313(273)	541(434)
					172(159)	273(238)	466(377)
						265(224)	245(206)
							179(156)

Table 3.1**.** Differential expression of EST contigs during the *A. gambiae* lifecycle. Numbers represent differentially expressed TCLAG contigs in each pairwise comparison (ANOVA Tukey test, P-value ≤ 0.001). The number of contigs that display at least two-fold difference is shown in parentheses.

The Embryo-high programme, EH displays characteristic strong expression in embryos. Annotation information revealed the prevalence of 6 major functional classes of genes: those involved in replication and transcription (as identified by DNA binding domains), mRNA processing and regulation, the cell cycle and its regulation, signal transduction pathways, cell growth and metabolism.

Most contigs with DNA binding domains seem to encode proteins involved in transcriptional regulation. This list includes orthologues of the *D. melanogaster* genes *extradenticle* and *brahma*, which encode proteins with a general polymerase II activity and are expressed in unfertilised egg and embryos, respectively. It also includes the orthologue of the fruitfly gene *extramacrochaetae (emc)* which has a maternal effect on embryonic development (Bellotto et al., 2002) The emc protein is known to dimerise with Achaete scute, lethal of scute and daughterless to inhibit DNA binding and transcription (Cabrera et al., 1994). The presence of sequences encoding ubiquitin domains suggests the involvement of ubiquitination in regulating mosquito embryonic development (Daniel et al., 2004).

The EH programme also encompasses a variety of genes that encode RNA binding elements, putatively involved in mRNA splicing and downstream processing. This list includes orthologues of the *Drosophila* genes *nonA-like*, *Pabp2*, *snRNP70k, sans fille, U1af50, pUf68, Gbp*, *Rbp1-like, Psi*. In addition, it includes the orthologue of the *D. melanogaster* gene *cornichon,* which is involved in *bicoid* mRNA localisation and formation of the anterior-posterior and dorsal-ventral embryonic axes.

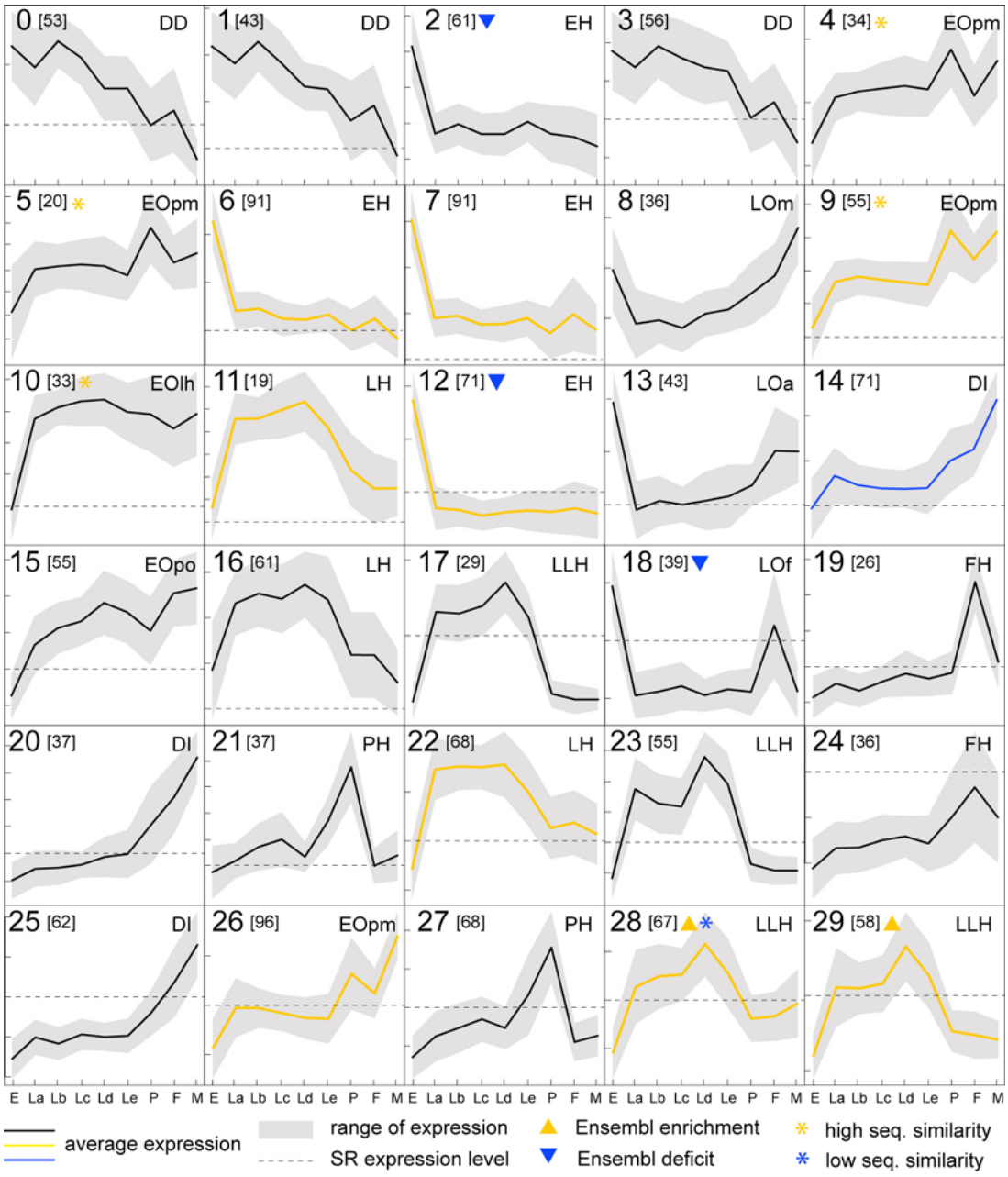


Figure 3.8. Developmental transcription programmes. During the mosquito lifecycle, 1,571 differentially expressed EST contigs that display at least two-fold expression differences between their maximum and minimum expression are grouped into 30 (0-29) SOM co-expression clusters. Numbers in parentheses indicate cluster size in contigs, solid lines indicate average expression and grey areas indicate range of expression. Each y-axis scale shows increments of 0.5 in log10-transformed expression values; horizontal dashed lines indicate SR signal levels. Arrowheads indicate clusters enriched (yellow) or deficient (blue) in Ensembl gene models. Average expression is shown in yellow or blue when genes in those clusters display average expression similarity with their Drosophila orthologues that is statistically above or below average relative to the median, respectively. Similarly, yellow or blue asterisks indicate clusters in which the genes have average coding sequence similarity with their Drosophila orthologues that is statistically above or below average, respectively. (please refer to developmental transcriptomic analysis section for more details).

The *embryo low programme, EO,* is defined by low embryonic expression and encompasses the subprogrammes: EOpm (high expression in pupae and adult males, suggesting possible expression in the testis (Belyakin et al., 2005)), EOpo (decreased pupal expression) and EOlh (broadly higher expression in larvae). Twenty-five percent of the EO contigs encode proteins involved in metabolic reactions, with the most prominent group being these involved in proton and electron transport. Several contigs correspond to putative components of immune responses: members of the clip-domain serine protease homologues, CLIPA1 and CLIPA10, the thioester-containing protein TEP7, the putative galactoside binding C-type lectin CTLGA3 and the leucine rich repeat immune gene *LRIM1* (Christophides et al., 2002; Osta et al., 2004). EO also encompasses genes encoding components of the cytoskeleton, including *gelsolin* (Vlachou et al., 2005) and the orthologue of *Drosophila Tropomyosin 1*, which plays a role in muscle formation and localisation of *oskar* mRNA in the embryo.

The *larva high programme, LH,* differs subtly from *EOlh* by showing a clear decline in Le; the expression is sustained at all earlier larval stages, when the imaginal discs that will give rise to adult structures grow slowly but the larval body grows faster. The latter feature is associated with high prevalence (ca. 30%) of sequences implicated in metabolic reactions (carbohydrate and lipid metabolism and proteolysis).

The *late larva high programme, LLH,* differs from LH in showing a clear peak of expression at Ld; it is highly enriched in sequences encoding metabolic enzymes (especially proteolytic), cuticle components. Contigs include putative immunity genes: members of the Gram negative binding protein subgroup B (GNBPB1, 2 and 4), C-type lectins (CTL3 and CTL4), the scavenger receptors SCRBQ1 and SCRB7, the serine protease inhibitor SRPN3, the melanisation inhibitor CLIPA2 (Volz et al., 2005) and the pro-phenoloxidase PPO3.

The *pupa high programme, PH,* is consistent with distinctive features of the pupal metamorphic period of holometabolous insects, when many larval tissues histolyse, while adult structures develop and often deposit adult cuticle. In mosquitoes, metamorphosis is also a transition from the aquatic feeding niche of the larva to the terrestrial life-style of the adult, including nectar feeding in both sexes and hematophagy uniquely in females. Indeed, the genes in this programme show peak expression in the pupa and correspond to structural and enzymatic components of the

cuticle, including members of the *yellow* family that functions in pigmentation in *D. melanogaster*. Other components are implicated in ubiquitination and proteolysis, suggesting a putative role in histolysis.

The *female high programme, FH* is enriched in genes encoding putative immune components, possibly suggesting adaptation to increase survival of the almost completely monogamous female mosquitoes. It includes a novel gram negative binding protein (GNBP), two clip-domain serine proteases (Shen et al., 2000), a fibrinogen-domain immunolectin, two products with allergen domains, a haeme peroxidase and three proteins with chitin-binding domains; one is ICHIT which is induced in mosquitoes both after bacterial challenge and malaria infection (Dimopoulos et al., 1998).

The *developmentally increasing programme, DI* displays a characteristic pattern of low expression in embryos and larvae and a dramatic increase in pupae and especially adults. After emergence from the pupal cuticle, many adult organs such as salivary glands, midgut epithelium and organs used for flight and orientation continue to develop. The maturation processes are known to begin earlier in males than in females, presumably explaining an observed difference in sex-specific expression (Clements and Clements, 1992). Some TCLAG encode products involved in adult physiological processes, e.g. the odorant proteins OBP10, OBP25 and OBP57 (Xu et al., 2003), members of the rhodopsin signalling pathway, the orthologue of the *D. melanogaster* flightin protein and others involved in digestion. Additional contigs encode antimicrobial peptides (DEF1, GAM and CEC3) or other immune-related proteins: the peptidogycan recognition protein, PGRPLB, a lysozyme orthologous to the *D. melanogaster Lysozyme D* (Kang et al., 1996), and a serine protease inhibitor, SRPN15.

The *developmentally declining programme, DD,* displays a characteristic pattern of overall progressive, albeit not continuous, decline in gene expression. Almost 20% of the expressed sequences encode products involved in protein biosynthesis, protein modification and folding; others correspond to DNA, RNA and nucleotide synthesis.

The *larva low programme, LO,* is characterised by high expression in both embryos and adults and is subdivided into LOa (equal expression in both adult sexes), LOm (higher expression in males) and LOf (higher expression in females). Most of the orthologues encode proteins with known expression in Drosophila embryos and adults. Examples are *exuperantia,* which is responsible for *bicoid*

localization in the embryo (Cha et al., 2001), *cactus, Pellino, inx6, Imitation SWI, Pendullin, maternal expression at 31B and moladietz,* the majority of which are found in LOf. Thus this cluster is enriched in genes that either have maternal effects on embryos or are involved in asymmetric mRNA or protein localisation.

Expression profiles of gene functional categories

Analysis of developmental expression by functional categories revealed that related GO terms or INTERPRO domains display similar expression profiles during the mosquito lifecycle (Fig. 3.9 and 3.S1- 3.S4). Biological processes related to the nuclear DNA replication, transcription and RNA processing activities display top expression in embryos and bottom expression in much of the remaining lifecycle (Fig. 3.9A1, 3.S2-0 and 3.S2-4) and mostly map to the EH transcriptional programme. Together these clusters highlight fundamental post-fertilisation processes: rapid succession of cell cycles associated with chromatin replication and initiation of transcription and translation for patterning the embryo and laying out the body plan. In contrast, processes related to general protein synthesis, folding and targeting have top expression in larvae and bottom expression in pupae and adults (Fig. 3.9A2); they map to the LH programme.

In contrast, a GO cluster displaying mostly bottom expression in embryos, larvae and pupae and top expression in adult males and females is enriched in processes related to adulthood such as vision and spermatogenesis (Fig. 3.9A3). Similarly, numerous domains implicated in immune reactions including antimicrobial peptides display bottom expression in embryos and top expression in pupae and adults (Fig. 3.9B1), indicating enhanced activity of the immune system after the larval phase. This cluster maps mainly to the DI and EO programmes, which contain many immunity genes. Whether high expression in adults reflects increased risk of infections of the terrestrial adults remains to be elucidated. Previous studies have documented robust expression of antimicrobial peptides in the adult midgut, which is reduced after administration of antibiotics (Richman et al., 1997). Additionally, several INTERPRO domains implicated in protein-protein interactions that serve diverse processes such as signalling and protein folding and ubiquitination tend to have higher pupal and adult expression (Fig. 3.9B2).

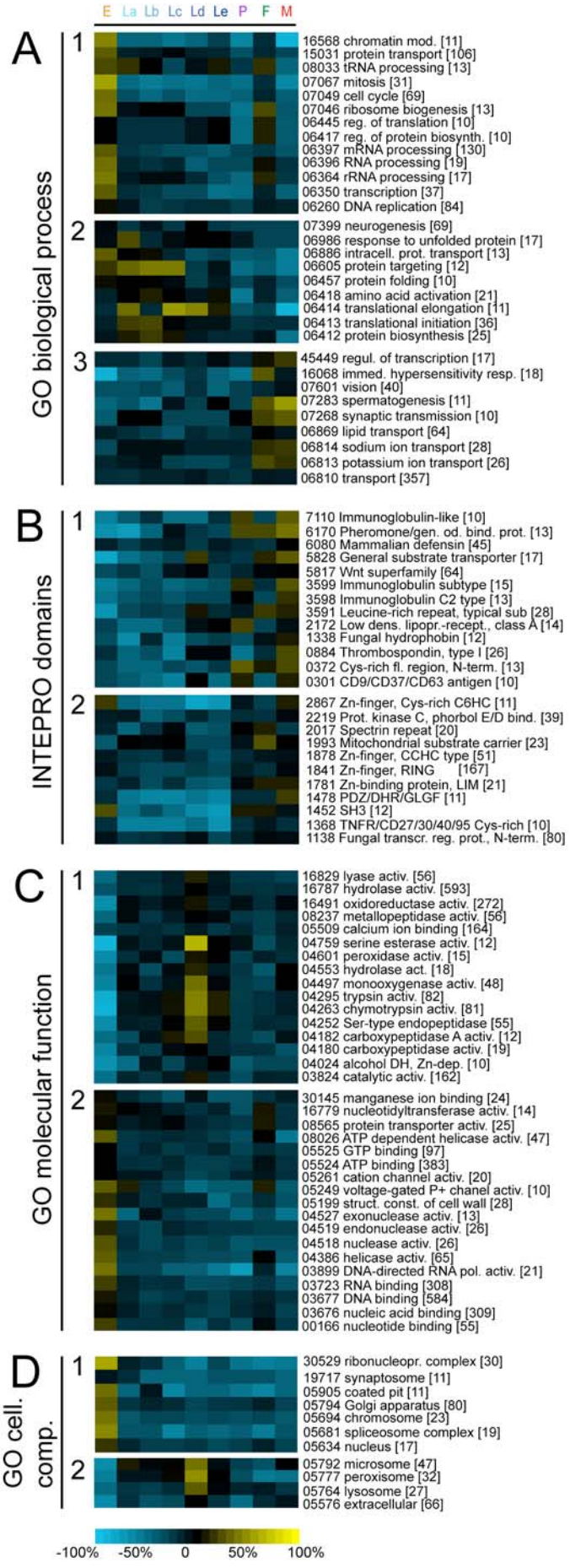

E La Lb Lc Ld Le P F M
A
GO biological process
1
16568 chromatin mod. [11]
15031 protein transport [106]
08033 tRNA processing [13]
07067 mitosis [31]
07049 cell cycle [69]
07046 ribosome biogenesis [13]
06445 reg. of translation [10]
06417 reg. of protein biosynth. [10]
06397 mRNA processing [130]
06396 RNA processing [19]
06364 rRNA processing [17]
06350 transcription [37]
06260 DNA replication [84]
2
07399 neurogenesis [69]
06986 response to unfolded protein [17]
06886 intracell. prot. transport [13]
06605 protein targeting [12]
06457 protein folding [10]
06418 amino acid activation [21]
06414 translational elongation [11]
06413 translational initiation [36]
06412 protein biosynthesis [25]
3
45449 regul. of transcription [17]
16068 immed. hypersensitivity resp. [18]
07601 vision [40]
07283 spermatogenesis [11]
07268 synaptic transmission [10]
06869 lipid transport [64]
06814 sodium ion transport [28]
06813 potassium ion transport [26]
06810 transport [357]
B
INTEPRO domains
1
7110 Immunoglobulin-like [10]
6170 Pheromone/gen. od. bind. prot. [13]
6080 Mammalian defensin [45]
5828 General substrate transporter [17]
5817 Wnt superfamily [64]
3599 Immunoglobulin subtype [15]
3598 Immunoglobulin C2 type [13]
3591 Leucine-rich repeat, typical sub [28]
2172 Low dens. lipopr.-recept., class A [14]
1338 Fungal hydrophobin [12]
0884 Thrombospondin, type I [26]
0372 Cys-rich fl. region, N-term. [13]
0301 CD9/CD37/CD63 antigen [10]
2
2867 Zn-finger, Cys-rich C6HC [11]
2219 Prot. kinase C, phorbol E/D bind. [39]
2017 Spectrin repeat [20]
1993 Mitochondrial substrate carrier [23]
1878 Zn-finger, CCHC type [51]
1841 Zn-finger, RING [167]
1781 Zn-binding protein, LIM [21]
1478 PDZ/DHR/GLGF [11]
1452 SH3 [12]
1368 TNFR/CD27/30/40/95 Cys-rich [10]
1138 Fungal transcr. reg. prot., N-term. [80]
C
GO molecular function
1
16829 lyase activ. [56]
16787 hydrolase activ. [593]
16491 oxidoreductase activ. [272]
08237 metallopeptidase activ. [56]
05509 calcium ion binding [164]
04759 serine esterase activ. [12]
04601 peroxidase activ. [15]
04553 hydrolase act. [18]
04497 monooxygenase activ. [48]
04295 trypsin activ. [82]
04263 chymotrypsin activ. [81]
04252 Ser-type endopeptidase [55]
04182 carboxypeptidase A activ. [12]
04180 carboxypeptidase activ. [19]
04024 alcohol DH, Zn-dep. [10]
03824 catalytic activ. [162]
2
30145 manganese ion binding [24]
16779 nucleotidyltransferase activ. [14]
08565 protein transporter activ. [25]
08026 ATP dependent helicase activ. [47]
05525 GTP binding [97]
05524 ATP binding [383]
05261 cation channel activ. [20]
05249 voltage-gated P+ chanel activ. [10]
05199 struct. const. of cell wall [28]
04527 exonuclease activ. [13]
04519 endonuclease activ. [26]
04518 nuclease activ. [26]
04386 helicase activ. [65]
03899 DNA-directed RNA pol. activ. [21]
03723 RNA binding [308]
03677 DNA binding [584]
03676 nucleic acid binding [309]
00166 nucleotide binding [55]
D
GO cell. comp.
1
30529 ribonucleopr. complex [30]
19717 synaptosome [11]
05905 coated pit [11]
05794 Golgi apparatus [80]
05694 chromosome [23]
05681 spliceosome complex [19]
05634 nucleus [17]
2
05792 microsome [47]
05777 peroxisome [32]
05764 lysosome [27]
05576 extracellular [66]
-100% -50% 0 50% 100%

Figure 3.9 (previous page). Functional groups of clusters exhibiting similar top/bottom expression. Selected k-means clusters of GO biological processes (A), INTERPRO domains (B), GO molecular functions (C), and GO cellular components (D). For each functional group, the percentages of TCLAG contigs showing top or bottom expression (upper 25% and lower 25% of the expression range, respectively) at each time period were calculated. Percentage of contigs showing bottom expression were subtracted from the respective percentage of contigs showing top expression, and resulting values (ranging from blue to yellow) were used for the k-means clustering. Numbers on the left side denote functional group identifier and numbers in brackets indicate the size of each functional group in EST contigs. The full analysis is presented in Figures 3.S1-3.S4.

The GO molecular function analysis provided interesting insights about general functions during the mosquito lifecycle. A wide variety of catabolic reactions are associated with bottom expression in embryos, pupae and adults but top expression in the Ld larval period (Fig. 3.9C1). Interestingly, in this cluster catabolic functions are also associated with some immunity genes such as GNBPs (hydrolase activity) and CLIPs (trypsin and chymotrypsin activities). This suggests that immunity functions may have evolved in catabolic, gut associated components that were in persistent contact with gut biota. The catabolic clusters are largely connected with the LH and especially the LLH developmental programmes, which also encompass numerous immunity genes. It is highly probable that this functional cluster mainly serves the histolysis of late larval tissues that mark the onset of metamorphosis. Consistent with this interpretation, contigs in specific subcellular organelles such as microsomes, peroxisomes and lysosomes show a similar pattern of expression (Fig. 3.9D2). The bottom expression of these contigs in embryos is consistent with this stage being marked by the initiation of transcription, translation and anabolic reactions. Indeed, the Fig. 3.9C2 and 3.9D1 clusters show top embryo expression of TCLAG contigs mostly involved in DNA and RNA replication, energy metabolism and anabolic reactions.

Coexpression patterns in specific adult female tissues

The study of spatial expression patterns identified 898 TCLAG contigs with reproducible differential expression between at least two tissues (Figure 3.7, tukey pairwise comparisons in Table 3.2 and dataset 3.D3). Of these, 829 exceeded two-fold regulation and were subjected to k-means clustering which grouped them into 10 distinct co-expression patterns (Figure 4). The gene contents of these patterns are summarized below and in Table 3.S2; the complete expression data are available in dataset 3.D4.

	Carcass	Gut	Ovaries
Head	295 (271)	172 (156)	370 (322)
Carcass		226 (212)	332 (302)
Gut			177 (161)

Table 3.2. Differential expression of EST contigs in adult female tissues. Numbers represent differentially expressed TCLAG contigs in each pairwise comparison (ANOVA Tukey test, P-value ≤ 0.001). The number of contigs that display at least 2-fold difference is shown in parentheses.

Head-enriched patterns (0, 1). The insect head carries most of the major sensory organs, the vision centre and endocrine glands. These patterns encompass contigs belonging to three major functional categories: rhodopsins and visual perception, odorant binding proteins and pheromone-related proteins. One contig encodes a homologue of the *Drosophila* Allatostatin, an adult brain peptide that blocks the synthesis of the developmental juvenile hormone.

Midgut specific patterns (2, 3). The midgut is the primary organ for absorption of nutrients, synthesis and secretion of digestive enzymes and peritrophic membrane formation. In addition, it has an endocrine role and contributes to diuresis, for example after a blood meal, when the blood cells are concentrated before digestion. Annotation identifies one quarter of these TCLAG contigs as implicated in metabolic reactions. Four TCLAG contigs were previously associated with midgut specific expression (Shen et al., 2000; Shen and Jacobs-Lorena, 1997; Shen and Jacobs-Lorena, 1998; Zheng et al., 1995), and three others encode a domain implicated in vasoconstriction and antidiuresis.

Carcass specific pattern (4) and Head and carcass-enriched pattern (5) Those patterns are small and contain contigs with diverse functions. They include genes putatively involved in immune responses (*CLIPA1*, *CLIPA6*. *CTLGA3*, *TEP7* and a serine protease homologue).

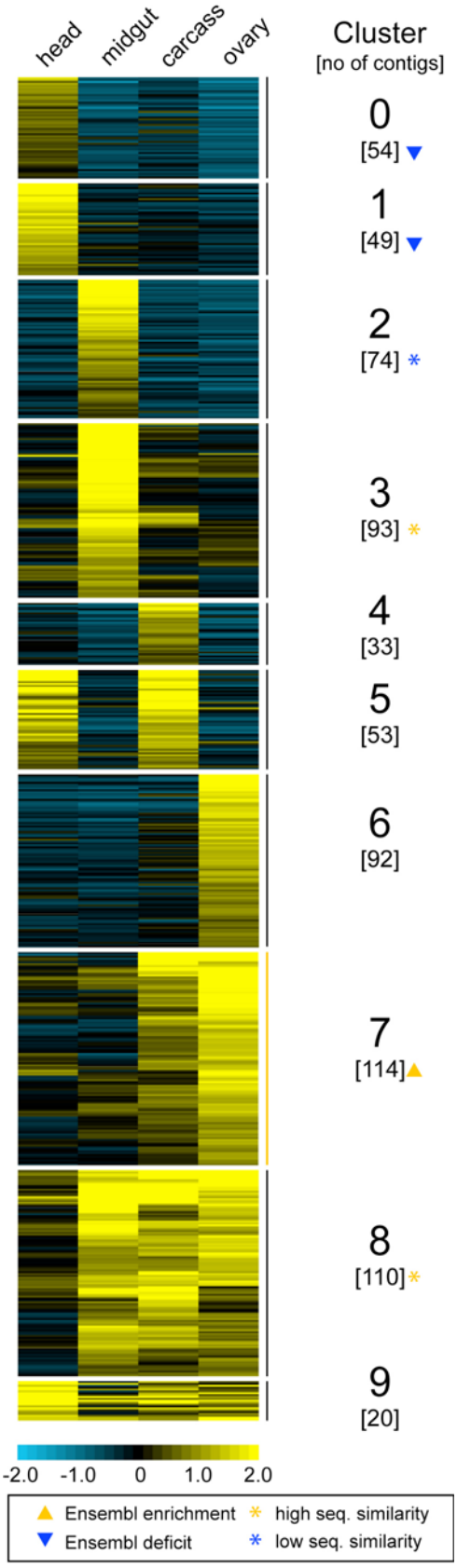

head
midgut
carcass
ovary
Cluster
[no of contigs]
0
[54]
1
[49]
2
[74]
3
[93]
4
[33]
5
[53]
6
[92]
7
[114]
8
[110]
9
[20]
-2.0
-1.0
0
1.0
2.0
Ensembl enrichment
Ensembl deficit
high seq. similarity
low seq. similarity

Figure 3.10 (previous page). Female tissue expression patterns. Adult tissue co-expression k-means clusters of 829 differentially expressed contigs, showing at least two-fold regulation between two tissues or more. Of these contigs, 477 correspond to Ensembl genes and 318 have *D. melanogaster* orthologues. Numbers in brackets indicate cluster size (in contigs) and scale bar represents gene expression (log_2 transformed) values. Yellow and blue arrowheads indicate clusters enriched or deficient in Ensembl gene models, respectively. Genes in the ovary-enriched cluster 7 (yellow sidebar) displays average expression similarity with their *Drosophila* orthologues that is statistically above average. Yellow or blue asterisks show clusters in which genes have average coding sequence similarity with their *Drosophila* orthologues that is statistically above or below average, respectively.

Ovary specific patterns (6, 7). The ovarian tissue is responsible for egg production after blood feeding in anautogenous mosquitoes. Those two co-expression clusters form a coherent ovary specific pattern but differ in respect of secondary expression in the carcass. This pattern is associated with the transcriptional programmes FH and LOf. One quarter is implicated in transcription regulation, translation and mRNA processing. Many other genes in this cluster are orthologues of *D. melanogaster* ovary specific genes such as *Cyclin A, Cyclin B, CyclB3* and *cdc2c* which are involved in cell cycle progression and genes such as *nanos, Pendullin, cactus, pole hole, spitz* and *vasa*, which have important roles in developmental processes and embryonic body patterning. Unexpectedly, this pattern also contains some members of the odorant binding protein family suggesting unorthodox functions that merit further analysis.

Midgut, carcass and ovary-enriched pattern (8). This pattern, of triple tissue origin is strongly enriched in metabolic functions. The annotation suggests general housekeeping processes: 20% of the encoded products are implicated in general metabolic reactions (polysaccharide and fatty acid synthesis) and 16% are involved in protein synthesis and degradation.

Four body part pattern (9). It differs from 8 in showing pronounced expression in the head. It contains diverse classes of molecules e.g. involved in electron transport, polysaccharide metabolism, signal transduction, ossification, proton transport etc.

Comparative transcriptomic analysis of Anopheles *and* Drosophila *lifecycles*

A previous comparative analysis of the *Anopheles* and *Drosophila* genomes (Zdobnov et al., 2002) has revealed remarkable sequence similarities, with more than half of the genes being 1:1 orthologues. The availability of a large-scale transcriptional study of *Drosophila* development (Arbeitman et al., 2002) allowed us to conduct a transcriptomic analysis of the lifecycles of the two insects. This was

achieved by comparing the developmental expression profiles of orthologous genes after normalisation of the two different experimental designs to create comparable notional developmental phases in the two studies (Table 3.3).

Notional Periods	***Anopheles***	***Drosophila***
Embryo	E	E056-E0112
Early larva	La-Lc	L24-L57
Mid larva	Lb-Ld	L43-L84
Late larva	Lc-Le	L67-L105
Pupae	P	M04-M12
Adult female	F	Af24
Adult male	M	Am24

Table 3.3. Comparable notional time periods in the *Anopheles* and *Drosophila* studies. The expression profiles were averaged when more than one time points were used.

Pearson and smooth correlation coefficients were used to assess the potential similarity of expression of orthologous gene pairs. Indeed, we detected a drastic shift towards positive correlation of the expression of orthologous genes by both methods (Fig. 3.11). In contrast when the same dataset was randomly rearranged 100 times to generate non-orthologous gene pairs, no shift was detected; both average distributions were largely symmetric. A clear, although not as pronounced, shift was detected using the non-parametric Spearman coefficient (data not shown). This analysis established that orthologues tend to share similar expression properties during the lifecycle of the two 250 mya-diverged insects.

This strong correlation suggested that orthologous genes may share common regulatory expression mechanisms that could potentially be reflected in sequence features of their promoters. Several algorithms for determining regulatory regions have been described (Tompa et al., 2005; Wasserman and Sandelin, 2004) which are mostly based on alignments of co-expressed sequences within a species. Our attempts to define a measure of similarity (Park et al., 2002) between 500 bp promoter regions of orthologous genes in the fruitfly, mosquito and honeybee were hindered by the inability to identify common evolutionarily conserved elements by a phylogenetic footprinting method (Blanchette and Tompa, 2002). Most likely, the predicting algorithms have not been trained to identify sequences conservation within such divergent datasets; comparisons between only three species are inherently unable to detect whatever promoter conservation remains after million of years of divergence.

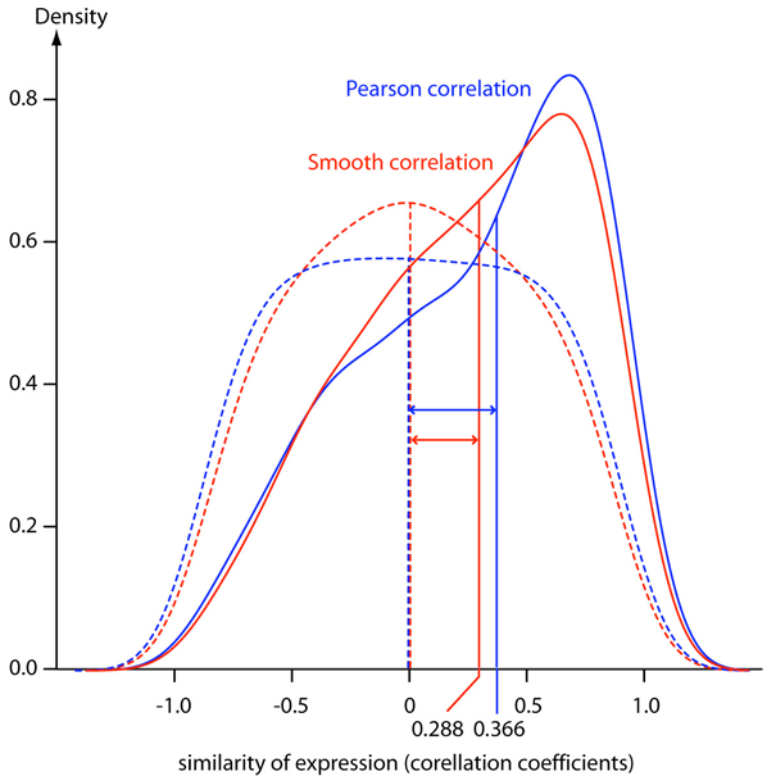


Figure 3.11. Positive correlation of orthologous gene expression between *Anopheles* and *Drosophila*. The distribution of orthologous gene pairs reveals a significant positive shift of expression correlation with both the Pearson (median/skewness = 0.366/-0.489) and the smooth (median/skewness = 0.288/-0.416) correlation coefficients. Dashed lines indicate the distribution of randomised non-orthologous pairs with the Pearson (median/ skewness = -0.003/0.005) and smooth (blue dotted line - median/skewness = -0.002/0.008) correlation coefficients.

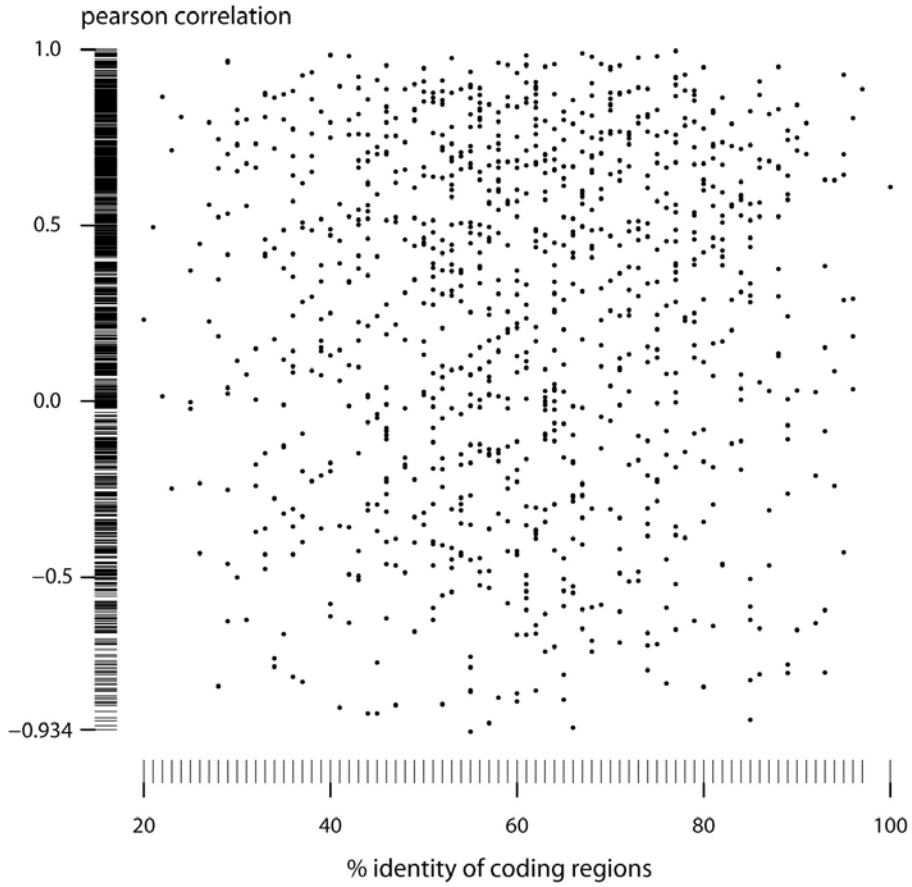


Figure 3.12. Scatter plot of similarity of expression (y-axis) versus the percent identity (x-axis) of the protein coding regions. Lines on each axis were added to show the density of each corresponding distribution. Note that no correlation between expression and sequence similarity between the 1,039 orthologous pairs can be established. (Both x and y-axes have been modified to show the density of the respective datasets).

We then examined whether the degree of interspecies correlation between developmental expression profiles of orthologous genes varies in parallel with the degree of coding sequence similarity. Surprisingly, no such global connection was detected (Fig 3.12).

Next, we queried whether a significant correlation between orthologous sequence similarity and expression similarity (according to the Pearson coefficient) might be detectable in smaller sets of genes, such as those engaged in particular developmental programmes or tissue patterns. As shown in Fig. 3.13, ten SOM co-expression clusters belonging to the EH, EOpm, DI, LH, and LLH programmes plus the ovary enriched pattern 7 showed significant deviations (positive or negative; Wilcoxon U-test ≤ 0.05) from the mean expression similarity, which was 0.259. Further, five SOM clusters belonging to the EOpm, EOlh and LLH programmes, plus the midgut-specific specific pattern 3 and the triple-bodypart pattern 8 showed significant

deviations from the median sequence similarity (Supplementary tables 3.S3- 3.S8). However, the overlap between these two deviating sets was limited to only 2 out of 16 clusters, whose deviations were not coherent: both developmental co-expression clusters 9 and 28 showed positive Δskewness in sequence similarity but negative Δskewness in expression similarity. We conclude that different evolutionary pressures affect sequence conservation of orthologous genes and the expression conservation.

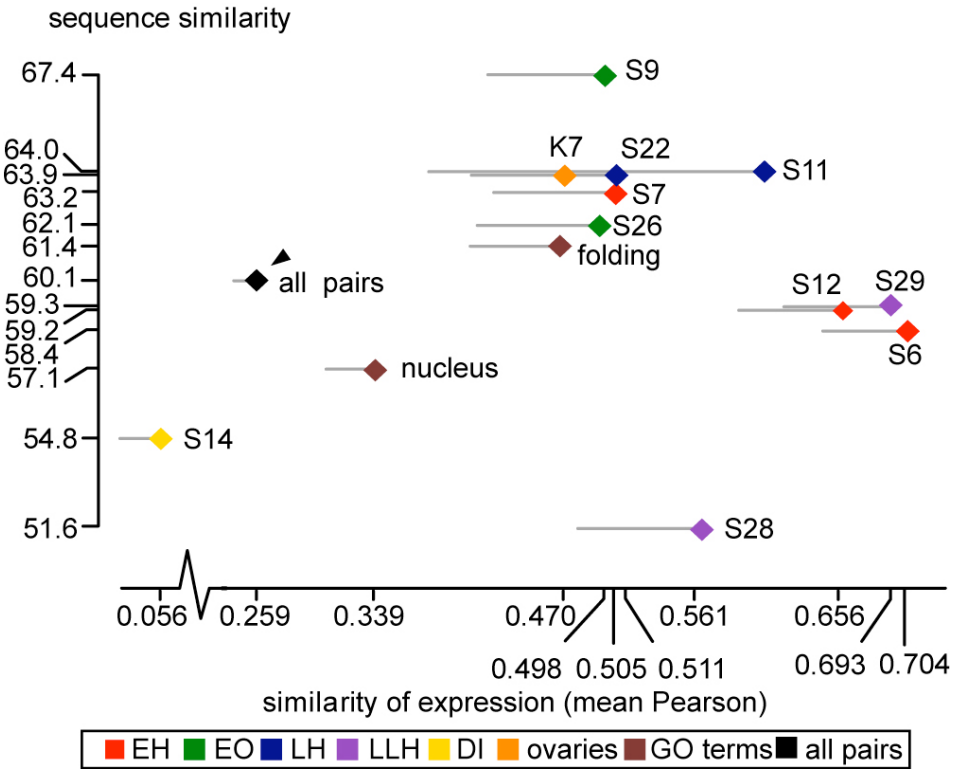


Figure 3.13. Scatter plot of the mean coding sequence similarity (y-axis) versus the mean expression similarity (x-axis) according to the Pearson correlation coefficient for developmental (S, SOMs) and tissue (K, k-means) co-expression clusters and GO functional groups. Only clusters and functional groups with significant deviation from the average expression similarity are presented. Coloured symbols refer to specific developmental programmes, functional groups or tissue patterns. A black arrowhead indicates the average sequence and expression similarity of all 1,039 orthologous gene pairs and grey lines show the standard error of expression similarity, plotted only on one side. The entire dataset is presented in Tables 3.S3- 3.S8.

Similarly, we searched for significant correlation between the levels of sequence and expression conservation in orthologues associated with specific functional groups (INTERPRO domains or GO categories). Out of a total of 42 interrogated functional groups, only two showed significant deviation from the median expression similarity. In particular, we detected significantly enhanced expression similarity in orthologues

encoding nuclear proteins, which often regulate gene expression, consistent with their enrichment in the high expression similarity EH programme. We also detected very high expression similarity for orthologues with protein folding function, consistent with their prevalence in the high expression similarity LH programme. On the other hand, 15 out of the 42 interrogated groups presented statistically significant deviation from the median sequence similarity of orthologues. Importantly, only one (the nuclear component group) showed significant deviation from both the median sequence and expression similarity (Fig. 3.13). However, even in this case the deviations were in opposite directions (higher than average expression similarity but lower than average sequence similarity), further supporting our conclusion that sequence and expression similarities of orthologues evolve independently.

Discussion

Anopheline mosquitoes have received scant attention despite their importance as vectors of devastating diseases and as model systems for the study of parasitism. However, determination of the genome sequence of *A. gambiae* in 2002 (Holt et al., 2002) opened unprecedented opportunities to dissect in molecular terms the biology of this major African malaria vector. Automated gene annotation analysis predicted approximately 14,500 genes in the *A. gambiae* genome, and comparative genomics highlighted their phylogenetic relations with genes identified in other sequenced genomes (Zdobnov et al., 2002). High-throughput transcriptional studies facilitated functional understanding of these genes, but have focused to date on specialised physiological processes such as immune responses (Christophides et al., 2002; Dimopoulos et al., 2002), responses to pathogens (Christophides et al., 2002; Dimopoulos et al., 2002; Kumar et al., 2003; Sim et al., 2005; Vlachou et al., 2005), haematophagy (Dana et al., 2005; Marinotti et al., 2006; Marinotti et al., 2005) and insecticide resistance (David et al., 2005; Vontas et al., 2005). Using strict quality criteria, the present study has profiled the transcriptional expression of the genome of *A. gambiae* throughout lifecycle and in several adult female tissues. By characterising the temporal and spatial co-expression properties of gene sets, rather than focusing on individual genes, it contributes a broad molecular understanding of the biology of this disease vector.

Anopheles developmental programmes and adult tissue patterns

A surprisingly limited set of nine temporal (developmental) expression programmes and six sub programmes, encompassing distinctive gene categories, appear to underpin the mosquito lifecycle. The *embryo-high* programme shows very characteristic temporal features, reflecting a rapid and specific activation of numerous genes that are implicated in embryonic development following fertilization. The EST contigs found in this programme are associated predominantly with processes related to DNA replication and transcription, post-transcriptional mRNA regulation and processing, and regulatory and signalling processes related to cell cycle and growth. In contrast, the *larva-high* developmental programme shows only minor differences

between different instars, possibly reflecting the continuous growth with increasing cell division rate that marks the larval period. The late larval period (*late larva-high* programme) is characterised by top expression of many metabolic enzymes that appear to usher metamorphosis. At the metamorphic transition from larva to adult, the *pupa-high* programme engages prominently genes that seem to be implicated in the synthesis of adult structures such as the adult cuticle. The greater similarity of gene expression between pupae and adult males (EOpm sub programme) indicates that the progression of adult development is faster in males. The *female-high* programme is enriched in immunity-related genes, suggesting increased immune activity and possibly competence of female mosquitoes. Females also share a number of similar expression profiles with embryos, recalling genes with known maternal effect in other species. Otherwise, male and female adult mosquitoes display only minor differences in gene expression.

The adult females of *A. gambiae* are highly important, as they transmit human pathogens including malaria parasites, filarial worms and viruses. Their transcriptomic analysis may identify gene sets that are related to the uptake of and susceptibility to these pathogens, ultimately facilitating the development of novel gene-based intervention strategies towards disease control. In this perspective, we identified spatial expression patterns for three critical female mosquito body parts, the head, midgut and ovary, which proved to include some very distinctive gene sets (clusters 0 and 1, 2 and 3, and 6 respectively). Predictably, the distinctive component of the carcass pattern proved to be rather small. The head contains not only the central nervous system but also most of the sensory organs that are responsible for host tracking and preference in female mosquitoes. Indeed, *A. gambiae* is such an important vector of human malaria primarily because of behavioural traits: the females prefer to feed on human rather than animal blood (anthropophily) and to nest and feed inside human habitations (endophily). Genes implicated in vision and odour-sensing are abundantly expressed in the mosquito head, and our analysis identified a plethora of such genes, many of which do not correspond to automatically predicted gene models. The midgut is the main organ for digestion of the blood meal and represents an important barrier for ingested human pathogens (reviewed in (Meister et al., 2004)). It also produces enzymes involved in a variety of metabolic processes. A previous transcriptomic study that used the same microarray platform has

examined genomic regulation in midgut cells during invasion by malaria parasites (Vlachou et al., 2005). The combination of these studies illuminates in considerable detail a large set of specifically midgut-expressed genes of the mosquito which might relate to permissiveness to malaria.

The ovaries, which two days after the blood meal contain mature eggs, have an expression pattern with a strong distinctive component. This pattern overlaps with the developmental programmes encountered in embryos, being enriched in components implicated in the basic gene expression processes (transcription, RNA processing and translation) as well as cell cycle control and patterning of the body plan. The carcass, corresponding to the mosquito remnant after removal of the tissues discussed above (and also legs and wings) shows expression profiles that are very diverse in terms of functional annotation and correspond to gene products involved in various metabolic processes; many are likely to be expressed in the fat body, the dominant tissue type of the carcass, which is equivalent to the mammalian liver. Future functional studies based on these tissue specific patterns are expected to identify genes that are critical for parasite transmission and will also permit analysis of regulatory regions implicated in these patterns leading to the discovery of tissue specific regulatory elements.

Microarray validation of previously identified EST contigs

We have previously reported that numerous EST contigs which are represented in MMC1 microarrays do not overlap with existing gene models (Kriventseva et al., 2005). The present study demonstrates that a high proportion of such contigs display specific developmental and adult tissue expression profiles, thus supporting our previous implication that many of them may correspond to true protein or RNA encoding genes which are missed by the automated prediction pipelines. Specifically, one third (505 of 1,571) of the EST contigs showing statistically significant differential expression in our analysis do not correspond to Ensembl gene models. Interestingly, the adult *head-enriched* pattern shows a pronounced deficit of gene models, as do the developmental programmes *embryo-high* and *larva-low-female-high* (presumably corresponding to embryonic and maternally inherited, embryonally-translated transcripts) (Figures 3.8 and 3.10). In contrast, the *late larva-*

high programme and the *ovary-enriched* pattern show a better than average coverage of transcripts by gene models. Taken together, these observations suggest that certain genes, which are highly expressed in embryos or encode putative maternal transcripts (in addition to those that are expressed strongly in the adult head), tend to be missed by automatic prediction algorithms. This seems a paradox as embryonic developmental expression is characterized very well in *D. melanogaster* and other insects; it suggests that many novel genes implicated in early development remain to be discovered by further studies of embryonic expression in *A. gambiae* and potentially other model systems. In clear contrast, late larval-specific (metamorphosis-associated) or ovary-enriched (as opposed to embryo-associated) genes have a high probability of automated prediction.

Expression similarity of orthologous genes between Anopheles *and* Drosophila

The recent availability of large expression datasets from several species has allowed interspecies comparisons of expression profiles of orthologous genes. Comparison of the rice and *Arabidopsis* transcriptomes showed above-average similarity of orthologous gene expression (Ma et al., 2005), especially for genes that are active in light as compared to dark conditions (Jiao et al., 2005). Similarly, conserved expression patterns of orthologues were identified in human and mouse adenocarcinoma cells (Stearman et al., 2005). Another study reported that genes belonging to specific GO categories share similar adult expression programmes in *Caenorhabditis elegans* and *D. melanogaster* (McCarroll et al., 2004). In contrast, a comparison of larval and adult profiles between the ant *Camponotus festinantus* and *D. melanogaster* revealed little expression similarity of orthologues (Goodisman et al., 2005). That study, however, utilised different data analysis procedures for each species dataset and only addressed a limited number of genes.

Several studies have addressed the issue of gene expression comparisons between microarray platforms. Early reports suggested little reproducibility between platforms (Kuo et al., 2002) or among widely used commercial platforms (Tan et al., 2003). Recently more systematic approaches addressed the issue of reproducibility between platforms and laboratories (Bammler et al., 2005; Irizarry et al., 2005; Larkin et al., 2005) and concluded that factors such as the use of pure starting material, biological

rather than abstract endpoints and similar data analysis procedures are likely to influence the analysis and determine the extent of agreement between different microarray platforms (Larkin et al., 2005). Here, having as guidance the above methodological conclusions, we performed a comparative analysis of the *A. gambiae* and *D. melanogaster* lifecycles utilising a dataset from a previously reported *Drosophila* developmental study (Arbeitman et al., 2002) and our new *Anopheles* study. To facilitate valid comparisons both datasets were analysed from the raw data, utilising the same quality criteria and normalisation procedures. It should be noted that *D. melanogaster* and *A. gambiae* display slightly different developmental stages: *Drosophila* encompasses three instar larval stages, as compared to four in *Anopheles*. Thus, instead of comparing biological endpoints, we based our analysis on notional time periods that were constructed based on aligning the temporal progression rather than fixed stages in the two species. This is the first large-scale comparative transcriptomic analysis of insect development.

The results reveal a strong positive correlation of expression throughout the lifecycle, for 1,039 orthologous gene pairs. The average similarity coefficients are comparable in magnitude to those reported in a previous adult rather than developmental interspecies comparison, of *D. melanogaster* and *C. elegans* (McCarroll et al., 2004). Importantly, we show that the degree of expression similarity is not globally linked to the degree of similarity in coding sequences, suggesting a non-concerted evolution of expression regulation and coding sequence in orthologues during the 250 million years of the dipteran clade.

Mosquito genes implicated in the EH, LH and LLH programmes display expression similarities with their fruitfly orthologues that are significantly higher than the average for all analysed orthologues. Taken together, these observations suggest that the broad biological similarity of embryonic and larval stages of the fruitfly and the mosquito is based in part on specific categories of orthologous genes that conserve their developmental expression profiles. The same may be true for the development of the ovary, as suggested by the higher than average expression similarity of orthologues belonging to the ovarian co-expression clusters. High expression similarity of orthologues is also encountered in genes implicated in the EOpm programme which shows low expression in embryos but high in pupae and adult males; provisionally we attribute this case of conserved expression to adult

male differentiation, possibly the development of the male gonads. Interestingly, genes implicated in the *developmentally increasing* programme, which predominantly reflects adult gene expression, display statistically significant divergence of expression from their *Drosophila* orthologues. This feature may imply strong diversifying selection pressure on expression exerted on these genes or their transcriptional regulatory elements, resulting in the known great physiological and ecological differences between adult mosquitoes and fruitflies.

Some gene sets that are defined by temporal or spatial co-regulation, or by functional classification schemes, may often also show a distinctive level of sequence similarity. We have noted that genes of the EOpm programme have a significantly high sequence similarity with their orthologous fruitfly counterparts strongly supporting high conservation of male differentiation between the two species. Coincident high conservation of both coding sequence and gene expression appears to be very rare indicating that these gene features evolve independently.

Conclusions

In conclusion, our study highlights specific properties of dipteran development. During the early developmental periods, genes involved in the processes of egg fertilisation, embryo formation and patterning of the body plan display conserved expression profiles. Later, as development ensues, differences in gene activation and expression become substantial and may account for the life cycle differences of the two insects, leading to adult organisms with very distinct lifestyles.

The developmental and tissue expression data that we report will be invaluable for understanding the biology of *A. gambiae* and for isolating regulatory regions that will permit engineered expression of transgenes in a temporal or spatial specific manner, towards future novel approaches to control vector populations. In addition, the present report represents a genomic scale comparative study of developmental evolution in a well-established model system (*D. melanogaster*) and another that is emerging (*A. gambiae*). Beyond the value for specialists, it contributes some concepts that may prove of broad relevance and interest, such as the apparent global dissociation of evolution between coding sequence and expression. This study also points towards a new concept for the correspondence between genes in different species: alongside sequence-based orthology, expression-based orthoregulation.

Chapter 3 Supplementary material

Programme	SOM clusters	TCLAG contigs	Ensembl genes	*Drosophila* orthologues
EH	2, 6, 7, 12	314	181	150
EO	4, 5, 9, 10, 15, 26	293	201	167
LH	11, 16, 22	148	114	89
LLH	17, 23, 28, 29	209	162	79
PH	21, 27	105	76	57
FH	19, 24	62	46	20
DI	14, 20, 25	170	114	84
DD	0, 1, 3	152	106	91
LO	8, 13, 18	118	65	46
Total		1571	1065	783

Table 3.S1**.** The *Anopheles* developmental programmes. Numbers represent the total amount of distinct TCLAG contigs, Ensembl genes and Drosophila orthologues that are included in the SOM clusters of each developmental programme.

Pattern	K-means clusters	TCLAG contigs	Ensembl genes	*Drosophila* orthologues
Head	0, 1	103	44	28
Midgut	2, 3	167	122	74
Carcass	4	33	27	12
Head-carcass	5	53	33	14
Ovary	6, 7	206	158	114
Midgut, carcass, ovary	8	110	80	63
Four body part	9	20	13	13
Total		692	477	318

Table 3.S2**.** The *Anopheles* adult tissue patterns. Numbers represent the total amount of distinct TCLAG contigs, Ensembl genes and Drosophila orthologues that are included in the k-means clusters of each adult tissue pattern.

Cluster	# of genes	Sequence similarity			Expression similarity		
		Δmedian	Δskewness	P-value	Δmedian	Δskewness	P-value
0	13	5	-0.183	0.266	0.157	-0.103	0.201
1	14	10	-0.280	0.111	0.253	-0.097	0.065
2	12	5.5	-0.979	0.534	0.218	-0.258	0.268
3	13	10	-0.747	0.109	-0.226	0.145	0.369
4	12	13	-0.964	0.025*	0.173	-0.648	0.546
5	6	17.5	1.547	0.003*	-0.024	0.794	0.363
6	17	-6	0.366	0.529	0.439	-1.700	<0.001*
7	25	5	-0.059	0.364	0.347	-0.680	0.004*
8	7	-3	1.590	0.797	-0.271	0.930	0.181
9	22	6	0.101	0.046*	0.240	-0.824	0.021*
10	9	19	-0.698	0.026*	0.331	-0.524	0.274
11	5	6	-1.122	0.490	0.490	-0.978	0.048*
12	11	1	-0.073	0.863	0.417	-1.040	0.004*
13	6	-8	0.278	0.243	-0.330	0.510	0.317
14	16	-7.5	0.243	0.206	-0.310	0.335	0.012*
15	8	0	0.277	0.956	-0.034	0.222	0.824
16	13	13	-0.598	0.264	0.159	0.060	0.176
17	2	7.5	0.026	0.490	0.063	0.489	0.552
18	7	4	-0.503	0.578	-0.161	0.898	0.666
19	2	-5	0.026	0.562	-0.059	0.489	0.780
20	4	-3.5	1.171	0.973	0.299	-0.040	0.164
21	4	-0.5	1.031	0.621	-0.208	0.339	0.420
22	18	1	0.040	0.362	0.268	-0.867	0.026*
23	2	5	0.026	0.609	0.241	0.489	0.322
24	6	-12.5	-0.255	0.060	0.156	0.558	0.340
25	10	-8	0.058	0.104	0.031	0.061	0.620
26	17	-1	0.058	0.662	0.351	-0.226	0.037*
27	12	6.5	-1.024	0.229	-0.004	0.036	0.924
28	14	-11.5	0.432	0.046*	0.268	-0.465	0.020*
29	7	0	0.022	0.872	0.333	-0.142	0.015*

Table 3.S3. Mosquito developmental programmes and coding sequence and expression similarities of *Anopheles – Drosophila* orthologues. For each cluster, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

Cluster	# of genes	sequence similarity			expression similarity		
		Δmedian	Δskewness	p-value	Δmedian	Δskewness	P-value
0	6	-6	-0.302	0.181	0.042	-0.200	0.996
1	3	7	-0.530	0.544	0.13	-0.189	0.899
2	9	-11	0.256	0.038*	0.076	0.201	0.611
3	18	7	-0.010	0.028*	0.045	0.128	0.556
4	6	4.5	-0.746	0.660	-0.178	0.590	0.706
5	11	-9	0.270	0.061	-0.369	0.020	0.166
6	11	-7	-0.520	0.052	0.215	-0.336	0.146
7	32	7.5	-0.149	0.240	0.149	-0.529	0.016*
8	22	15.5	-0.534	<0.001*	0.055	-0.024	0.903
9	7	10	-0.108	0.419	0.435	0.016	0.054

Table 3.S4. Mosquito tissue patterns and coding sequence and expression similarities of *Anopheles* – *Drosophila* orthologues. For each cluster, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

Functional Group	# of genes	sequence similarity			expression similarity		
		Δmedian	Δskewness	P-value	Δmedian	Δskewness	P-value
GO:0006810	82	3	-0.134	0.112	0.059	0.074	0.869
GO:0006355	65	-2	-0.158	0.128	0.098	0.013	0.534
GO:0008152	56	5	-0.208	0.032*	0.024	-0.009	0.484
GO:0007275	42	-2	0.148	0.150	0.063	-0.123	0.378
GO:0006412	41	8	-0.570	0.001*	-0.221	0.215	0.218
GO:0006118	31	9	0.099	0.009*	-0.065	0.351	0.699
GO:0015031	31	4	-0.114	0.125	-0.113	0.229	0.645
GO:0000398	28	-5.5	0.196	0.213	-0.115	0.234	0.402
GO:0006397	25	2	-0.220	0.993	-0.262	0.289	0.213
GO:0006457	25	-2	0.111	0.788	0.188	-0.714	0.044*
GO:0006508	24	-2.5	0.001	0.695	0.017	-0.018	0.930
GO:0007155	23	-7	0.551	0.104	-0.024	-0.146	0.811
GO:0006468	22	1	0.012	0.721	-0.096	0.267	0.116

Table 3.S5. GO biological process terms and coding sequence and expression similarities of *Anopheles* – *Drosophila* orthologues. For each GO term, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

Functional Group	# of genes	sequence similarity			expression similarity		
		Δmedian	Δskewness	p-value	Δmedian	Δskewness	P-value
GO:0005634	163	-3	0.035	0.039*	0.129	-0.205	0.032*
GO:0016021	137	2	-0.009	0.238	-0.018	0.095	0.736
GO:0016020	97	3	0.040	0.128	0.048	-0.117	0.825
GO:0005739	70	13.5	-0.890	<0.001*	0.023	-0.039	0.411
GO:0005622	35	7	-0.323	0.051	-0.083	-0.019	0.526
GO:0005783	30	4	0.418	0.074	-0.049	-0.046	0.923
GO:0005794	23	3	0.215	0.258	-0.254	0.415	0.095
GO:0005840	21	18	-1.117	<0.001*	-0.222	0.411	0.184

Table 3.S6. GO cellular component terms and coding sequence and expression similarities of *Anopheles – Drosophila* orthologues. For each GO term, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

Functional Group	# of genes	sequence similarity			expression similarity		
		Δmedian	Δskewness	p-value	Δmedian	Δskewness	P-value
GO:0016787	102	0	0.255	0.548	0.005	-0.130	0.756
GO:0005524	90	8	-0.313	<0.001*	0.090	-0.085	0.809
GO:0003677	80	-6	0.132	<0.001*	0.141	-0.243	0.058
GO:0016740	78	2	-0.328	0.107	-0.004	0.084	0.169
GO:0016491	67	5	-0.473	0.015*	0.026	-0.025	0.362
GO:0003723	50	-0.5	0.092	0.435	0.029	-0.108	0.700
GO:0003824	50	7	-0.292	0.012*	-0.038	0.068	0.955
GO:0003676	43	-5	-0.006	0.268	0.145	-0.350	0.176
GO:0008270	37	-8	0.272	0.036*	-0.019	0.284	0.244
GO:0005515	34	0	0.143	0.410	0.091	-0.235	0.352
GO:0005198	25	-5	0.236	0.282	-0.025	0.113	0.305
GO:0005509	25	-1	-0.324	0.807	-0.058	0.097	0.409
GO:0003700	23	-7	0.347	0.029	-0.193	0.382	0.439
GO:0004672	23	1	0.044	0.751	-0.061	0.197	0.183
GO:0005215	23	4	0.481	0.155	0.043	0.330	0.319
GO:0003735	22	18	-1.039	<0.001*	-0.230	0.447	0.150
GO:0005525	22	22.5	-0.994	<0.001*	0.083	-0.481	0.513
GO:0016829	22	7	-0.221	0.136	-0.096	0.408	0.851
GO:0016874	22	10.5	-0.366	0.002*	0.255	-0.329	0.313
GO:0004674	21	1	0.059	0.453	0.012	0.141	0.297
GO:0004842	20	-4	0.293	0.710	0.254	-0.203	0.365

Table 3.S7. GO molecular function terms and coding sequence and expression similarities of Anopheles – Drosophila orthologues. For each GO term, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

Functional Group	# of genes	sequence similarity			expression similarity		
		Δmedian	Δskewness	p-value	Δmedian	Δskewness	P-value
IPR008211	46	-1	0.057	0.537	0.145	-0.282	0.474
IPR001841	38	1.5	0.095	0.498	0.167	-0.574	0.156
IPR001138	22	1	-0.107	0.209	-0.064	-0.203	0.872
IPR003593	21	17	-0.139	<0.001*	0.075	0.137	0.906
IPR001007	20	1.5	-0.324	0.785	0.023	-0.034	0.707

Table 3.S8. INTERPRO domain and coding sequence and expression similarities of *Anopheles* – *Drosophila* orthologues. For each INTERPRO domain, we have calculated the difference in the median and the skewness of either the sequence similarity (coding sequence identity) or the expression similarity (Pearson coefficient). The Wilcoxon test (U-test) was used to test significant deviations from the 1,039 orthologous gene pairs. Asterisks denote gene groups with U-test P-values ≤ 0.05.

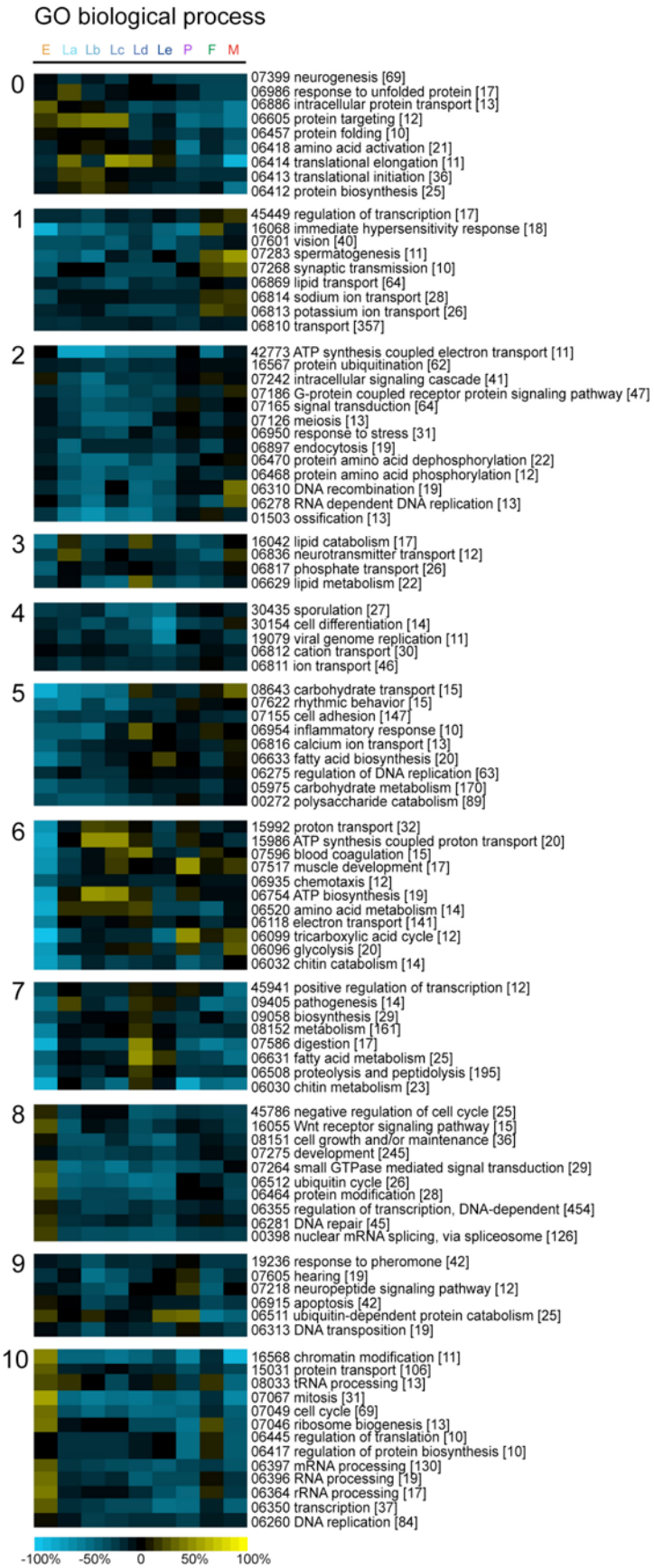


Figure 3.S1. GO biological process groups exhibiting similar top/bottom expression. Percentage of contigs in each functional group ranges from more bottom than top expression (negative values, blue) to more top than bottom expression (positive values, yellow). Numbers in brackets indicate the number of TCLAG contigs in each group and plain numbers indicate the GO identifiers.

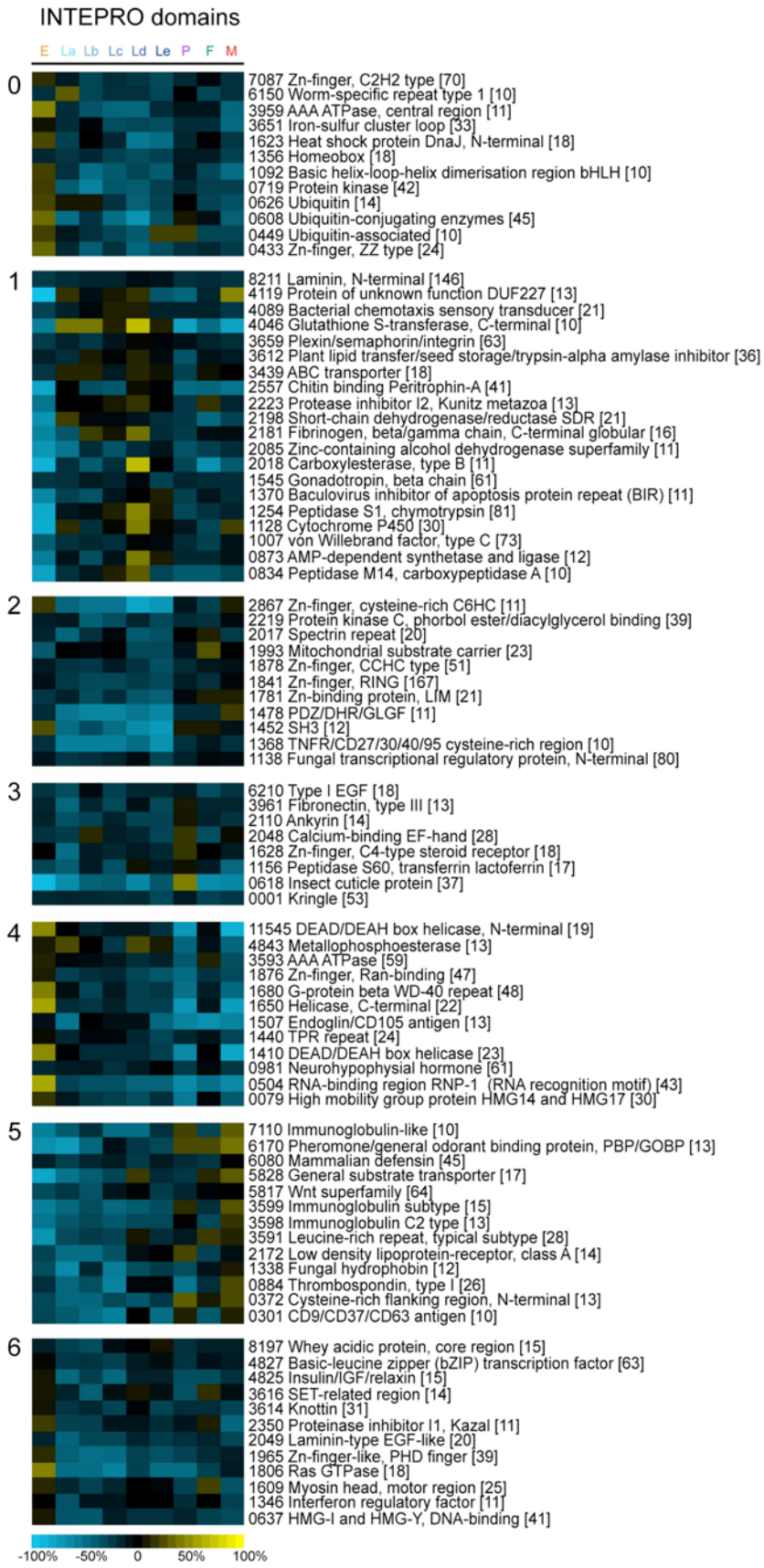


Figure 3.S2. INTERPRO domains exhibiting similar top/bottom expression. Percentage of contigs in each functional group ranges from more bottom than top expression (negative values, blue) to more top than bottom expression (positive values, yellow). Numbers in brackets indicate the number of TCLAG contigs in each group and plain numbers indicate the INTERPRO identifiers.

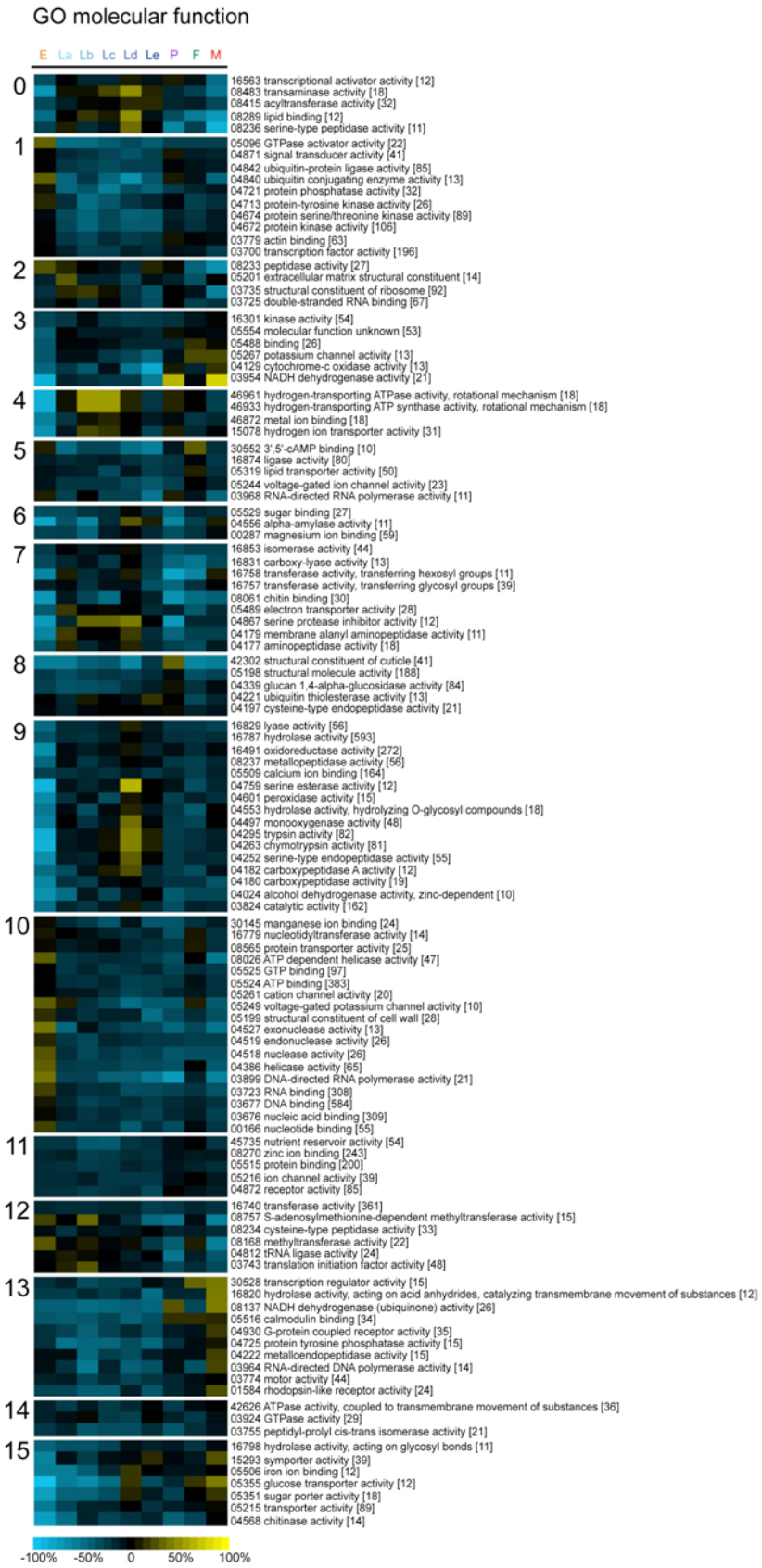


Figure 3.S3. GO molecular function groups exhibiting similar top/bottom expression. Percentage of contigs in each functional group ranges from more bottom than top expression (negative values, blue) to more top than bottom expression (positive values, yellow). Numbers in brackets indicate the number of TCLAG contigs in each group and plain numbers indicate the GO identifiers.

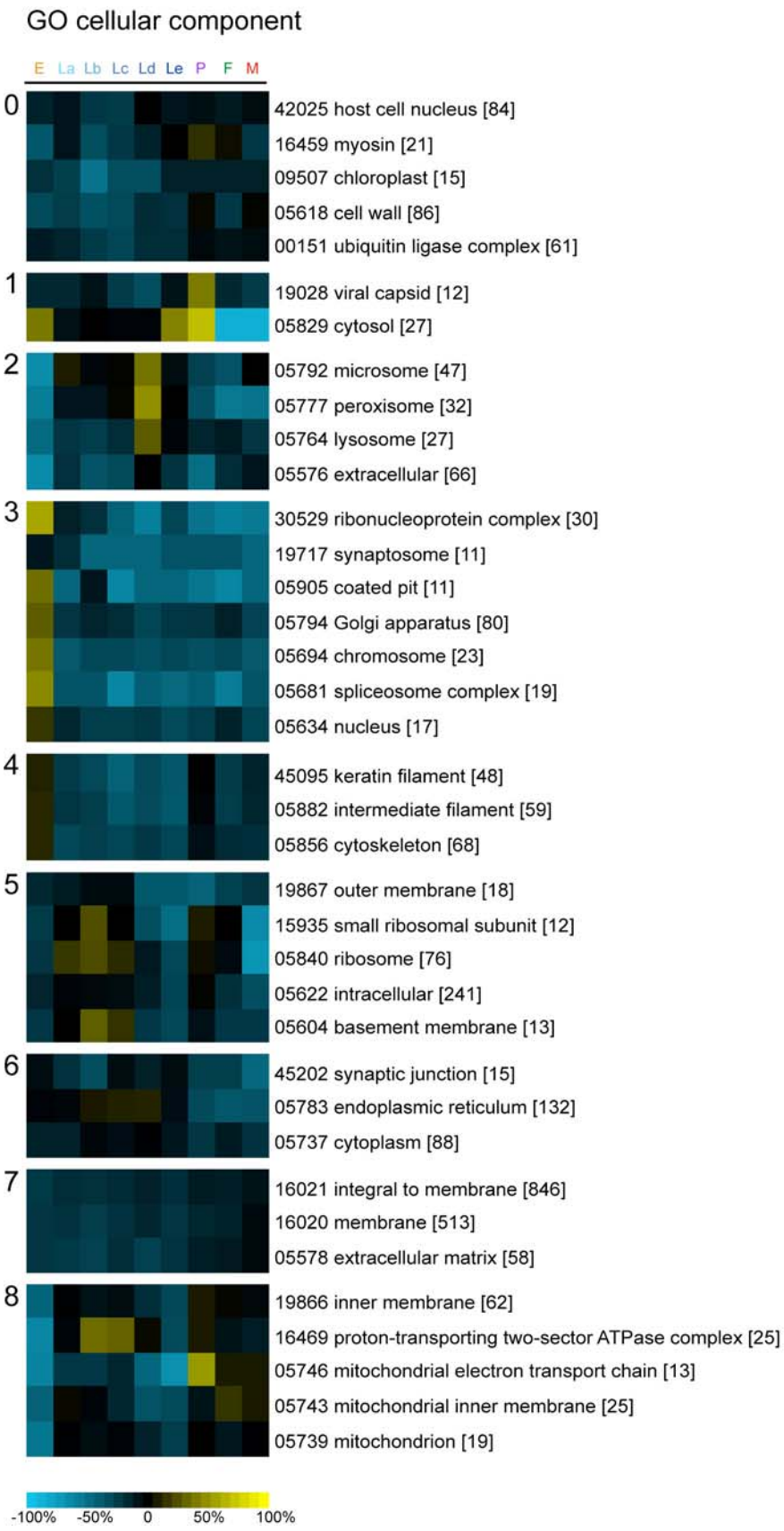
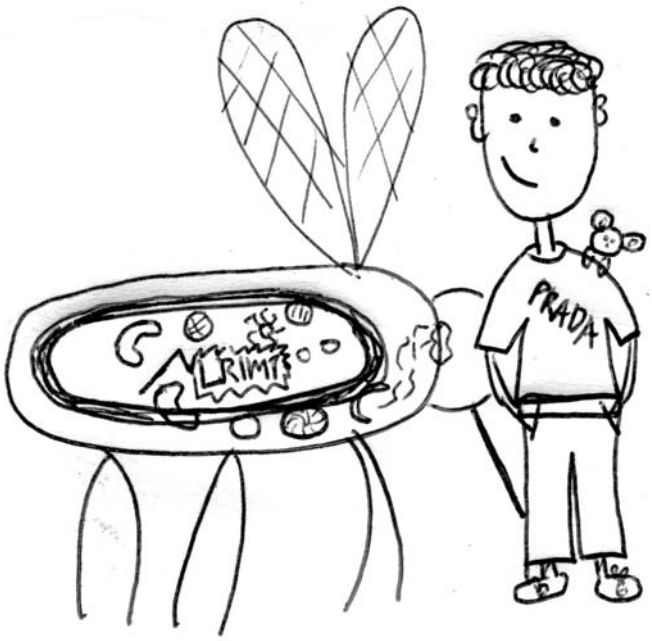


Figure 3.S4. GO cellular component groups exhibiting similar top/bottom expression. Percentage of contigs in each functional group ranges from more bottom than top expression (negative values, blue) to more top than bottom expression (positive values, yellow). Numbers in brackets indicate the number of TCLAG contigs in each group and plain numbers indicate the GO identifiers.



Chapter 4

LRIM1, a novel leucine rich repeat gene involved in innate immune responses against bacteria and malaria parasites

Major losses in the parasite phase in the mosquito

During the development of the malaria parasites in the mosquito vector several losses occur and the parasite population undergoes 'bottlenecks', which can reduce its size to single digit numbers. As evidenced from Fig. 4.1, two major bottlenecks in the mosquito phase occur which coincide with the transitions from the ookinete to the oocyst and from the midgut to the salivary glands sporozoites. Both these transition steps are characterised by parasite crossing of barriers (the peritrophic membrane and the midgut epithelium in the former case and the salivary gland epithelium in the latter).The first bottleneck is a major one and may lend to complete blockade of the infection. Several factors could explain the dramatic reduction in parasite number and one of them is a robust immune response mounted by the mosquito. If they survive those bottlenecks, the parasites will establish and infection at the salivary glands and – upon a subsequent mosquito bite – infect new hosts ensuring the survival and persistence of malaria.

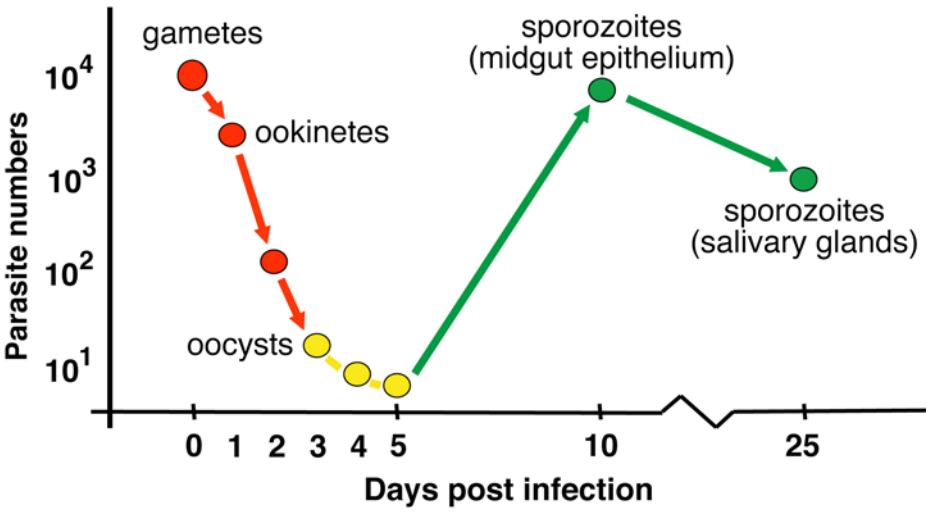


Figure 4.1. Major parasite losses in the mosquito. Two characteristic parasite bottlenecks in the mosquito are observed during the ookinete to oocyst and the midgut to salivary gland sporozoites transition. Conversely, major parasite amplification occurs in the oocyst stage. The initial number of ingested parasites is estimated to 10^4. (Figure courtesy of Blandin, S. and Levashina, E., modified from (Sinden, 1999)).

An overview of innate and adaptive immune responses

Organisms have developed mechanism to protect them from attacks from pathogenic microorganisms that are collectively called immune systems. There are mainly two types of immune systems; both the innate and the adaptive immune have co-evolved to allow organisms to identify and eliminate pathogens. The innate immune system provides the first line of defence by detecting the immediate presence and nature of infection. It relies on a limited number of receptors recognising invariant molecules that are shed by microorganisms. These immune inducers are referred to as pathogen associated molecular patters (PAMPs), including peptidoglycans (PGN), lipopolysaccharide (LPS), β-1,3-glucans, as well as glycosyl phospatidyl inositol (GPI). They are also found in non-pathogenic microorganisms.

A different strategy of immunity, called adaptive immunity was later discovered in evolutionary history and is based on the generation of a diverse and large number of recombinant proteins (B-cell and T-cell antigen receptors) created through genomic rearrangements. Innate and adaptive immunity have many differences: innate immunity is immediate and is shared by many organisms whereas adaptive is delayed and is exclusive to vertebrates. However, apart from its role in defence, innate immunity serves to trigger and direct the adaptive immune responses, as well as to gain time for the adaptive immune system to unfold its full effectiveness (Fearon and Locksley, 1996; Schnare et al., 2001). The main difference is that memory or specialisation to cope with infective agents relies solely in evolution for innate and to the additional strategies of clonal selection and proliferation for the adaptive.

The insect's first barrier against intruders is provided by the structural barriers of the body, including the hardened exoskeleton, the chitinous trachea and the peritrophic membrane of the midgut. Breaches in the first two are quickly sealed by coagulation (Theopold et al., 2002) and melanisation (Soderhall and Cerenius, 1998). Similarly the peritrophic membrane has among its functions a role in restricting microorganisms developing in the gut lumen (Shao et al., 2001). Beneath those barriers lie the respective epithelia, bathed in hemolymph, the insect blood. Epithelia serve as physical barriers can also mount strong responses against microorganisms (Ferrandon et al., 1998; Onfelt Tingvall et al., 2001; Tzou et al., 2000). Finally, should microorganisms invade the body cavity they encounter the robust immune reactions of the fat body and the blood cells.

In the last decades, we have experienced a dramatic increase in the knowledge of innate immune reactions, due to the studies in a variety of organisms that includes crabs, crayfish, ascidians and a variety of insect species (reviewed in (Iwanaga and Lee, 2005)). Among the insect examples, pioneering studies in *D. melanogaster* contibuted to the detailed dissection of the innate immune pathways and showed that the underlying mechanisms have been conserved during evolution (Hoffmann and Reichhart, 2002). Innate immune responses can be separated into two kinds: the humoral response with the production of antimicrobial peptides (AMPs) and melanisation and the cellular responses that include phagocytosis and encapsulation of intruders.

The pioneering studies in Drosophila have concentrated largely in the characterisation the components of two conserved immune pathways, the Toll and the Imd (Fig. 4.2), which are utilised to respond primarily to bacterial and fungal infections (Hoffmann, 2003). The Toll pathway was originally implicated in the dorso-ventral pattern formation during *Drosophila* development (Anderson et al., 1985; Hashimoto et al., 1988). Its central role in defence reactions against Gram-positive bacteria and fungi was later established, when loss-of-function mutation to this protein rendered flies vulnerable to infection by fungi (Lemaitre et al., 1996). Gram-negative bacterial infections on the other hand, are predominantly dealt with by activation of the Imd (Immune deficiency) pathway, named after the first identified mutation in the pathway (Lemaitre et al., 1995). However, recent studies have point out that there is not clear-cut assignment of group of pathogens to the immune pathway; for example, instead of the *Toll* mediated response, several fungi are activating the Imd pathway (Hoffmann, 2003; Hultmark, 2003).

The Toll *and the* Imd *pathways in Drosophila*

The first step to the activation of the Toll pathway is the recognition of PAMPs by specific receptor proteins (Janeway and Medzhitov, 2002). PGRP-SA (Michel et al., 2001) and gram negative binding protein, GNBP1 are involved in the recognition of PGN of Gram-positive bacterial, whereas GNBP3 is involved in the recognition of fungi (Ferrandon et al., 2004). Those proteins act in concert to mediate the signal to a proteolytic cascade, which has not been entirely deciphered. Studies have implicated the gene products of *Gastrulation defective, Snake* and *Easter* in Toll embryonic

activation and *Persephone* and *necrotic* in immune activation by fungi. The end result of this cascade in the activation and dimerisation of Spaetzle, the only factor that mediates binding to the Toll receptor (Lemaitre et al., 1996). The activation results in the intracellular recruitment of at least three cytoplasmic proteins, MyD88, Tube and Pelle. Pelle is a serine-threonine kinase believed to play an indirect role in the phosphorylation and subsequent proteolytic degradation of Cactus, a member of the I-κB family of proteins that normally bind Dorsal and Dif preventing their nuclear translocation. Depending on the developmental stage, degradation of Cactus and subsequent activation of Dorsal or Dif leads to their translocation in the nucleus. Dif is mostly implicated in the transcription of AMPs through specific binding to *cis*-acting elements (NF-κB) found in their promoter sequences.

Likewise, two PGN recognition proteins have thus far been shown to mediate activation of the Imd pathway: the transmembrane PGRP-LC (Choe et al., 2002; Gottar et al., 2002) and the extracellular PGRP-LE (Takehana et al., 2002). After an initial recognition event, the signal is passed on through an unknown process to the Imd, an adaptor protein carrying a domain, Death, commonly associated with proteins controlling apoptosis (Lemaitre et al., 1995). A series of downstream events involving the *Drosophila* homologs of the mammalian IκB kinase complex (IKK) (Silverman et al., 2000) and caspase Dredd lead to the proteolytic cleavage and activation of Relish (Stoven et al., 2000), the third member of the NF-κB family of transcription factors. Full length Relish contains an amino-terminal DNA binding domain (transcription factor) and a carboxy-terminal IκB domain, which acts similarly to the Toll pathway inhibitor Cactus, preventing the nuclear translocation of the transcription factor domain when the pathway is inactive (Dushay et al., 1996). After proteolytic removal of the IκB domain, Relish translocates into the nucleus and induces transcription of AMPs.

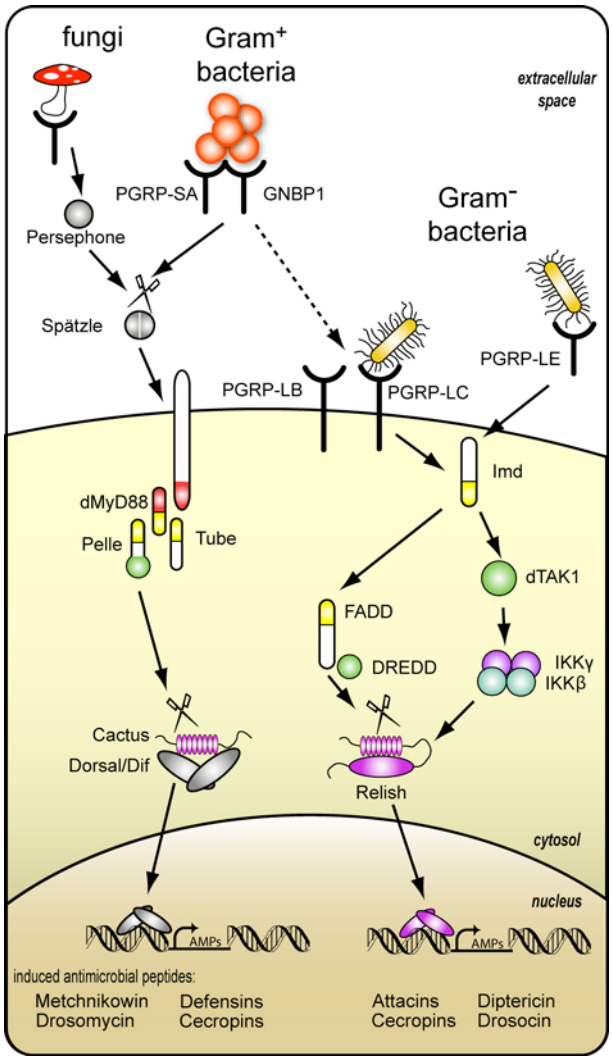


Figure 4.2. Schematic representation of the Toll and the Imd pathways in *Drosophila*. The *Toll* pathway is principally activated by fungi and Gram-positive bacteria and the Imd pathway by Gram-negative bacteria. The result from the activation of the each pathway is the translocation of the NF-κβ proteins Dif, Dorsal and Relish, which enter the nucleus and activate the synthesis of antimicrobial peptides. (fig. adapted from (Meister et al., 2004)).

The end result of the activation of both the Toll and the Imd pathway is the production of AMPs. These are typically small, cationic and structurally diverse peptides secreted into the hemolymph soon after infection (Bulet et al., 1999). Although the main source of AMPs is the fat body, various epithelia, the genital tract and the malpighian tubules are also capable of producing these peptides (Tzou et al., 2000). The exact mode of action varies between different AMP families; however, in general, AMPs kill microbes by attacking the structure of the cytoplasmic membrane, for example by permeabilisation or by forming voltage-dependent channels (Bulet et al., 1999).

There have been seven distinct families of AMPs identified in *Drosophila* that have various target specificities (Hetru et al., 2003). Defensins (Dimarcq et al., 1994) are the most widespread family of AMPs in insects and other invertebrates and act mainly against Gram-positive ($Gram^+$) bacteria whereas Cecropins (Kylsten et al., 1990), although having a broader spectrum, are more effective against Gram-negative ($Gram^-$) bacteria. Other families of AMPs include Attacin (Asling et al., 1995), Diptericin (Wicker et al., 1990), Drosocin (Bulet et al., 1993), Metchnikowin (Levashina et al., 1995) and Drosomycin (Fehlbaum et al., 1994), with the last being solely antifungal.

The melanisation reaction

Melanisation is a mechanism of wound healing, cuticle sclerotisation and humoral responses that entails the production of a melanin polymer as the result of the conversion of tyrosine to melanin. A link between the coagulation and prophenoloxidase (PPO) cascades leading to melanisation has been shown in insects and crustaceans (Li et al., 2002; Nagai and Kawabata, 2000). The PPO cascade is triggered by recognition of microbial cell wall components such as PGN (Yoshida et al., 1996), LPS and β-1,3-glucan (Lee et al., 2000; Ma and Kanost, 2000). This leads to the activation of a serine protease cascade that culminates in the proteolytic cleavage of PPOs to form active POs. Negative regulation of the melanisation reaction is thought to be achieved through serine protease inhibitors (serpins, SRPNs), which act as suicide substrates of the PPO activating serine proteases (PPAEs) (De Gregorio et al., 2002; Ligoxygakis et al., 2002; Zhu et al., 2003). A proposed model in *Drosophila* is laid on a pre-activation balance between the inhibitory serpin and the PPAE, which changes in favor of the PPAE upon Toll pathway activation (Ligoxygakis et al., 2002). The final steps of this reaction are performed by phenoloxidases (POs) that catalyse the oxidation of tyrosine to dihydroxyphenylalanine (DOPA), and subsequently to dopaquinone and dopaminequinone. These quinones form the precursors of a melanin polymer that kill the invader by enclosing it in a proteinaceous melanin capsule.

Cellular reactions: phagocytosis and encapsulation

Cellular responses include phagocytosis and encapsulation and are mediated by circulating blood cells that are collectively called hemocytes. In *Drosophila* larvae, there are at least three classes of morphologically distinguishable hemocytes: plasmatocytes, lamellocytes and crystal cells (Meister and Lagueux, 2003). Phagocytosis is performed by plasmatocytes, the main population of hemocytes, while encapsulation is carried out by lamellocytes. Finally, crystal cells are thought to mediate melanisation of encapsulated bodies in a reaction called melanotic encapsulation.

Phagocytosis is the cellular process in which blood cells recognise, internalise and destroy microbial invaders (Aderem and Underhill, 1999). This process typically includes recognition of the microorganism destined for destruction, activation of intracellular cascades that lead to actin polymerisation, extension of filopodia and internalisation of the invader. Initial binding of opsonic ligands to the microorganism appears to be necessary for its recognition by the phagocytic cell. A receptor involved in the phagocytosis of Gram-negative bacteria in *Drosophila* is none other than the Imd pathway mediator PGRP-LC, indicating a link between humoral and cellular immune reactions (Ramet et al., 2002). Recently, a transmembrane protein with EGF like repeats, *Eater*, was reported to mediate direct binding to bacterial and be involved in phagocytosis (Kocks et al., 2005).

The Anopheles innate immunity

Until recently, not many details have been known about molecular pathways of immune responses in *Anopheles*. However, with the publication of the genome sequence, allowed an *in silico* genome wide approach for the identification of the major families of proteins involved in innate immunity that shared homology to the corresponding families in *Drosophila* (Christophides et al., 2002). This study, along with several functional studies (discussed later in this chapter) provided the basis for the elucidation of the *Anopheles* immune reactions in response to different pathogens.

The Toll and Imd pathways represent two of the most documented pathways of immune responses. In *Anopheles*, orthologues of PGRP-SA (*PGRPS1*) and the intracellular signalling molecules (*MYD, TUBE, PLL1*) were found in *Anopheles*; however, importantly, no orthologues of *Dif* and *GNBP1* were detected, and *Toll*

forms an orthologous group together with four mosquito genes (*TOLL1A, TOLL5A, TOLL1B, TOLL5B*) (Christophides et al., 2002; Luna et al., 2002). The conservation of the Imd pathway, however, is very likely since orthologues of all the aforementioned genes (*PGRPLC, IMD, IKK1, IKK2, REL2, CASPL1*) and other component (*TAK1, FADD,*), except for PGRP-LE, were identified in its genome (Christophides et al., 2002). However, notable differences have been observed; whereas *Relish* responds only to Gram-negative bacteria in *Drosophila*, *REL2* in responds to both Gram-positive (*Staphylococcus aureus*) and Gram-negative (*Escherichia coli*) bacteria (Meister et al., 2005).

Four families of AMPs have been identified so far in *Anopheles*: Defensins (*DEF*), Cecropins (*CEC*), Attacin and Gambicin (*GAM1)* (Christophides et al., 2002; Vizioli et al., 2001a; Vizioli et al., 2001b; Zheng and Zheng, 2002). Consistent with other Defensins, the mosquito *DEF1* is most active against Gram-positive bacteria, yeast and filamentous fungi, but not against Gram-negative bacteria (Richman et al., 1996; Vizioli et al., 2001b). However, *CEC1*, unlike *Drosophila* Cecropins, is more active against yeast and a number of Gram-positive bacteria (Richman et al., 1996), supporting the hypothesis for functional divergence of immunity genes. Interestingly, the same gene is significantly upregulated in *Plasmodium*-infected mosquitoes (Christophides et al., 2002; Dimopoulos, 2003). Finally, *GAM1* belongs to a family identified so far only in *Anopheles* and *Aedes*, and has a broad antibacterial spectrum; it is also effective against *Plasmodium* (Vizioli et al., 2001b).

The Plasmodium parasite, an additional challenge for Anopheles innate immunity

The majority of studies that have formulated a firm understanding of the mechanisms of the immune responses in *Drosophila* have been carried out using specific types of bacterial pathogens and fungi. Studies to the nature and kind of immune responses to different organisms, mainly protozoans and viruses had been therefore lacking. In *Anopheles*, the *Plasmodium* parasite represents a 'new challenge' for the innate immune responses. A number of studies have shown that the malaria parasite activates the expression of several mosquito genes, implicating, at least in part, mosquito immune responses for the documented parasite losses (Dimopoulos et al., 2002; Dimopoulos et al., 1997; Dimopoulos et al., 1998; Richman et al., 1997; Tahar et al., 2002; Vlachou et al., 2005). Whether the mosquito

utilises the known innate immune factors of bacterial and fungal infection or has evolved new molecules for the recognition and destruction of the parasite, and whether those factors act in concert, is the subject of ongoing research and remains to be elucidated.

To date, the best examples of immune responses against parasites in the mosquito are the parasite melanisation and parasite killing, which have been observed in specific mosquito strains. Parasites are melanised almost immediately after crossing the midgut epithelium in a genetically selected *Anopheles gambiae* strain called L3-5 (Collins et al., 1986). As a result, this mosquito strain completely blocks the development of *P. cynomologi, P. berghei,* and some allopatric strains *of P. falciparum* and has been termed refractory. However, it does not melanise sympatric strains of *P. falciparum* (Collins and Paskewitz, 1995; Paskewitz et al., 1988), supporting the theory that the reaction involves a parasite specific recognition. The ability of these mosquitoes to melanise Sephadex beads (Paskewitz and Riehle, 1994) as well as the observation of bead melanisation in field-collected mosquitoes (Schwartz and Koella, 2002) supports the latter hypothesis. However, a recent report showed that a C-type lectin knockdown-induced melanisation is directly involved in killing of parasites, whereas melanisation which is encountered in refractory mosquitoes merely disposes dead parasites (Volz et al., 2006).

Studies for the genetic attribution of the melanisation phenotype in refractory mosquitoes have implicated three quantitative trait loci, called the Pen loci that accounted for approximately 70% of the variance of ookinete killing and melanisation (Zheng et al., 1997). The *Pen1* genomic region shows clusters of extensive sequence polymorphisms that may relate to the refractory phenotype (Thomasova et al., 2002). However, different loci were reported to control the melanisation of the *Plasmodium cynomolgi* parasite, suggesting that different loci are involved for different parasites (Zheng et al., 2003). Furthermore, a recent multidisciplinary morphological, biochemical and genomic approach has demonstrated broad physiological differences between the refractory (L3-5 strain) and susceptible (G3 strain) and detected an elevated level of reactive oxygen species as another factor that contributes to parasite melanotic encapsulation (Kumar et al., 2003). In addition to melanisation, another parasite killing mechanism in the mosquito midgut has been reported for an *A. gambiae* strain that eliminates *P. gallinaceum* ookinetes (Vernick et al., 1995).The ookinetes appear to be initially

vacuolated and subsequently lysed while still in the cytoplasm of the midgut epithelial cells. Henceforth, we will call this refractory mechanism parasite lysis.

An important tool towards the understanding of the role of individual genes in mosquito immunity and *Plasmodium* infection has been the development of gene specific silencing by RNA interference (RNAi). This phenomenon was initially observed in plants (Napoli et al., 1990; van der Krol et al., 1990) and was then referred to as posttranscriptional gene silencing but its wide application in other organism was later defined with the description of this phenomenon in *C. elegans* (Fire et al., 1998). Blandin et al demonstrated that injection of dsRNA into the thorax of adult mosquitoes results in the efficient and transient downregulation of the specific gene (Blandin et al., 2002). Since then, this technique has been widely employed to investigate the function of mosquito genes such as members of the thioester containing proteins (Blandin et al., 2004), serine protease inhibitors (Abraham et al., 2005; Michel et al., 2005), components of the Toll and Imd pathways (Meister et al., 2005), C-type lectins (Osta et al., 2004), clip-domain serine proteases (Volz et al., 2005), genes involved in phagocytosis (Moita et al., 2005) or putatively involved in actin cytoskeleton dynamics (Vlachou et al., 2005) and several other gene families (Kafatos, FC, unpublished results).

A gene that clearly plays a role in parasite lysis and melanisation is TEP1, a member of the family of thioester containing proteins. Knockdown (KD) of *TEP1* by RNAi silencing resulted in a dramatic increase in the number of developing oocysts in susceptible mosquitoes, as well as an increase of oocyst number and inhibition of melanisation in the refractory melanising mosquitoes (Blandin et al., 2004). TEP1 was previously implicated in phagocytosis of bacteria (Levashina et al., 2001) and was found to bind *P. berghei* ookinetes, mediating their lysis in the midgut cells (Blandin et al., 2004).

The discovery of a new gene family involved in mosquito immune responses

A previous large –scale transcriptomic analysis identified several genes that are differentially expressed after a variety of immune challenges in cell lines and adult mosquitoes (Dimopoulos et al., 2002). Among those genes, 4 candidates genes with leucine rich repeat (LRR) domain structure were upregulated in bacterial and parasite challenges (Fig 4.3). Several other lines of evidence suggested a role of these genes

in immune responses. Three members are upregulated after parasite midgut invasion (Vlachou et al., 2005) and another is regulated by *REL2* (Meister et al., 2005). A characteristic of these proteins is that they do not show significant homology to any other known proteins and their domains structure implicates them in general protein-protein interactions or pattern recognition. However, in the absence of any other functional studies, their role in innate immune reactions remained, until recently, largely hypothetical.

Leucine rich repeats domain structure and function

Members of the leucine rich repeat (LRR) family of proteins serve diverse functions; they are described as hormone receptors, enzyme subunits, enzyme inhibitors, cell adhesion proteins, ribosome binding proteins and immunity proteins (Kajava, 1998; Kobe and Deisenhofer, 1994; Kobe and Deisenhofer, 1995). Their common characteristic is a domain consisting of multiple tandem copies of a 20-29 amino acid (aa) sequence containing leucine residues. Sequence analyses suggested the classification of LRR proteins into at least 7 different subfamilies according to the length of the LRR residues and the type of their secondary structure (Kobe and Kajava, 2001). The significance of this classification is that repeats from different subfamilies never occur simultaneously in the same protein suggesting independent evolution. A typical example of an LRR motif consists of a conserved eleven-residue segment, LxxLxLxxN/CxL, which corresponds to the structure of a β-sheet and its adjacent loop segment and 13 additional residues, which may vary and usually correspond to a-helical secondary structures. The invariant part of the β-sheets is a common feature of the LRR motifs whereas the variant part suggests different functional adaptation and permits the classification of the proteins in the aforementioned subfamilies.

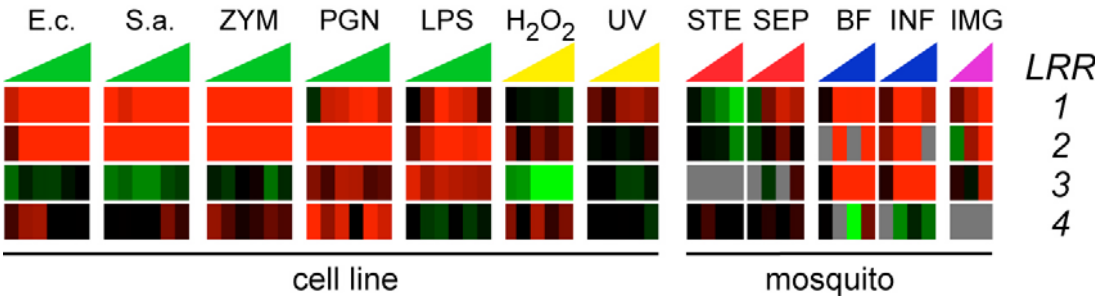


Figure 4.3. Upregulation of mosquito LRR containing proteins after immune challenges. Microarray expression transcriptomic profiling across 12-time course experiments in cell lines (left) and adult female mosquitoes (right). E.c: *Escherichia coli*, S.a.: *Staphylococcus aureus*, ZYM: zymocan, PGN: petidoglycan, LPS: lipopolysaccharide, UV: ultraviolet radiation, STE: sterile injury, SEP: septic injury (*E.coli*, *S. aureus*), BF: blood feeding, INF: infection, IMG: infected midgut. Red colour: upregulation, green colour: downregulation, black colour: no regulation, gray colour: no data.

The first solved structure of an LRR-containing protein was that of the porcine ribonuclease inhibitor (Kobe and Deisenhofer, 1993) and provided an insight to the arrangement of the LRRs in 3-dimensional space. In this model, the β-sheets and a-helices are arranged in parallel to a common axis, resulting in a characteristic, non-globular horse-shoe shaped molecule (Fig. 4.4). The β-sheets line the inner circumference creating a hydrophobic region in the solenoid structure, whereas the helices flank the outer circumference. The domain appears curved because the variant part of the motif is bigger than the invariant part of the β-sheets. Since then, the numerous solved LRR structures provided a framework for understanding the variability of the horse-shoe shaped structure (Fig. 4.5). Sequence and size differences of the variant part account for either an increase or a decrease in curvature, which is reflected in a wider or a narrower structure of the protein (compare the curvature of LRR-containing proteins of Fig. 4.5). Other atypical LRR-structures include the substitution of the a-helices with the characteristic 3_{10} domain (Marino et al., 1999) and the leucine-rich invariant repeats, in which a-helices line the inner circumference and 3_{10} domains the outer circumference (Andrade et al., 2001).

Figure 4.4. Three-dimensional structure of porcine ribonuclease inhibitor (PDB: 2BNH), the first solved LRR protein. In the characteristic solenoid, globular, horse-shoe shaped structure composed of the LRR repeats the a-helices (magenta) and β-sheets (yellow) of each repeat are parallel to a common axis. (Fig. created with default coloring scheme of secondary structures in RasMol v2.6 software).

The modular architecture of the LRRs functions for protein-protein interactions (Bell et al., 2003; Bergelson et al., 2001; Chamaillard et al., 2003; Vasselon and Detmers, 2002). The side-to-side association of repeats builds an arch, with the β-sheets forming the interior of the arch harboring an extensive protein-binding surface. Specificity to the interacting partner is conferred by a number of other additional factors: a) the number of LRR motifs, b) the residues of the loop segments of the β-sheets, c) deletions or insertions between the individual LRRs and d) additional residues at the C-terminal of the LRR-motif (Bell et al., 2003; Huizinga et al., 2002; Kajava, 1998).

Leucine rich repeats in immunity proteins

Invertebrates and mammals contain a variety of LRR containing molecules that are involved in innate immune reactions. The motifs are shared by the Tolls of *Drosophila* and other invertebrates and the Toll like receptors (TLRs) of vertebrates

and other organisms, in the NBS-LRR (NODs) intracellular proteins of mammals and the resistance (R) genes in plants.

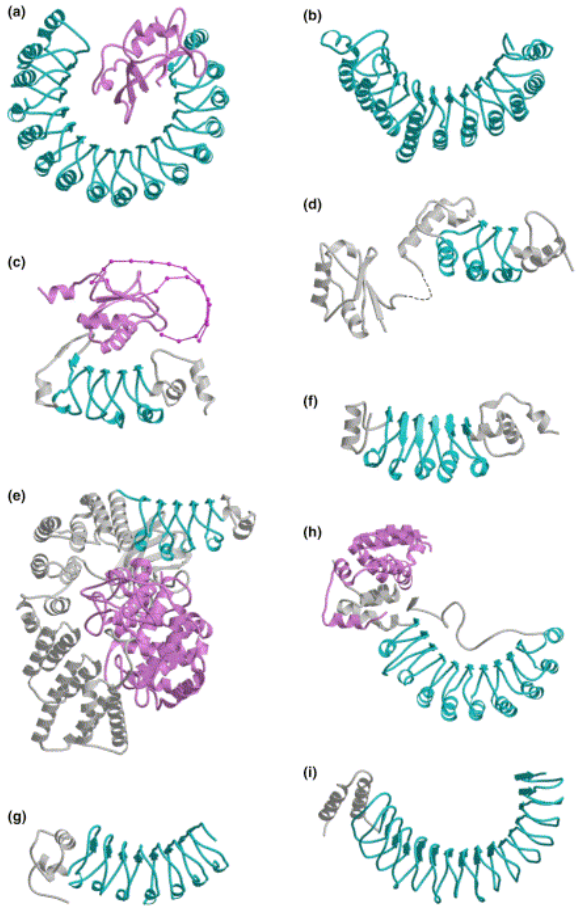


Figure 4.5. Three-dimensional structures of LRR proteins. LRR domains are depicted in cyan, flanking regions in grey and other parts in magenta. Notice the variability in the number of LRR repeats and the curvature of the horseshoe-shaped structure. (a) RI (PDB code 2BNH), b) rna1p (1YRG), c) U2A'-U2B'' (1A9N), d)TAP (1FO1), e) RabGGT (1DCE), f) dynein LC1 (1DS9), g) InLB (1DOB) h)Shp2-Skp1 (1FQV) i)YopM (1G9U). From (Kobe and Kajava, 2001).

The first member of the Toll and Toll-like receptor (TLR) family of LRR proteins was the *D. melanogaster* Toll protein. Since then, seven additional molecules with structural similarity to Toll have been detected in *Drosophila* but their involvement in innate immune reactions remains to be established. In mammals, several proteins have been found to share homology with the *Drosophila* Toll (Bell et al., 2003; McGuinness et al., 2003; Vasselon and Detmers, 2002). The structure of the Toll and TLR proteins consists of an extracellular LRR domain, a transmembrane domain that mediates the anchorage of the protein and an intracellular domain called Toll / IL-1 domain (TIR), which displays homology to the interleukin-1 element. The Toll family of receptors act as intermediate molecules in the transferring of the pathogen

signal, by linking the extracellular compartment, where the contact and recognition of the microbial pathogen occurs, to the intracellular space, where the signal is transferred though the activation of the TIR domain and its interaction with further downstream proteins. This recognition can be indirect, as in the case of Toll and direct, as in other TLR proteins.

Organisms have also developed a system for the intracellular recognition of pathogen molecules. Central to this system are the NBS-LRR family of proteins that have been recently reported (reviewed in (Chamaillard et al., 2003)). The characteristic of this family is the tripartite domain structure, which consists of a C-terminal LRR, a central nucleotide binding domain (NBS) and a protein interaction domain at the N-terminus, which divides the family into several different subfamilies. Members of the NBS-LRR family include the NOD1, NOD1 and Naip proteins, which have been in implicated in the intracellular recognition of PAMPs. NOD2 recognises muramyl dipeptide, a specific peptidoglycan motif from bacteria and activates an immune response (Girardin et al., 2003; Inohara et al., 2003). In addition, many autosomal immune diseases in mammals have been linked to mutations in the genes encoding NBS-LRR and associated proteins, suggesting a possible link between bacterial surveillance and autoimmune diseases.

In plants, the initial phase of immune responses involves the recognition of the PAMP by receptor-like LRR kinases and the activation of the signalling cascade that leads to the immune responses. Bacteria inactivate such immune responses by delivering effector molecules to the plant cytoplasm though a type-III secretion system. Ultimately, plants evolved the resistance genes (R) to identify the bacterial effector proteins and restore immune responses by activating the so-called effector-mediated transcription. R proteins contain predominately LRR domains and are subdivided into the intracellular NB-LRRs and the extracellular eLRRs. The former group as subdivided into two further subclasses, coiled-coil or TIR containing proteins and the latter family contains three subclasses: the receptor-like proteins with extracellular LRR and TM domains, the receptor-like kinases with extracellular LRR, TM and kinase domains and the polygalacturonase with a cell wall LRR. In R genes, the recognition of the PAMP is almost always indirectly. There are, however, LRR containg R-proteins which do not fall into any of the above categories, suggesting variable functions.

The above examples have established a central role of LRR proteins in innate immune responses. Recently, however, it has been shown that the LRR motif can also be employed in adaptive immune responses (Pancer et al., 2004). Jawless fish such as sea lampreys produce a large number of LRR-containing variable leukocyte receptors (VLR) via a recombination strategy similar to the antibody and T-cell receptor strategy in mammals. The receptors are encoded by a single genomic locus, gVLR, which contains information for the 5' and the 3' segments of the mature gene products but lack information about the intermediate region that is characterised by the presence of a variable number of LRR domains. Production of the mature VLR molecules occurs with the recombination of upstream of downstream LRR cassettes to the intermediate region of the gVLR locus. Thus, although the mechanism of production of variable molecules is conserved in vertebrates, different evolutionary choices may have contributed to the use of LRRs for the agnathan (jawless fish) and immunoglobulin molecules for the gnathostomes (jawed vertebrates).

Aims of the current study

The publication of the genome sequence of the malaria vector mosquito, *A. gambiae* paved the way for functional genomic studies. The development of DNA microarrays allowed for the large-scale assessment of gene expression and identification of genes whose transcription is altered after a variety of immune challenges. Among the identified candidates, 4 genes with LRR domains did not show any homology to known immune related proteins in other species, even though LRR proteins are found in several innate immune responses in *Drosophila*, mammals and plants.

Therefore, the aim of the current study is a more in-depth investigation of those genes and their involvement in mosquito immune reactions. This chapter reports results from the first candidate gene of this family, *Leucine rich repeat immune gene 1 (LRIM1)*. We have investigated the involvement of *LRIM1* in immune reactions against bacteria and its effect on *Plasmodium* development in mosquito midguts. RNA KD experiments of *LRIM1* display a phenotype similar to that observed for *TEP1*: a dramatic increase in the number of developing oocysts in the susceptible mosquito, as well as an increase in oocysts and abolishment of melanisation in

refractory mosquitoes. In addition, we present evidence for the localisation of LRIM1 during parasite midgut invasion.

Materials and Methods

In silico bioinformatic tools for the domain characterisation of LRIM1 protein and LRR domain modelling

Gene, transcript and protein identifiers for *LRIM1* are shown in table 4.1. For the domain characterisation of LRIM1, the protein sequence was subjected to a variety of *in silico* domain prediction programs that included SMART (Letunic et al., 2006; Schultz et al., 1998) for the general prediction of domains, signalP (Nielsen et al., 1997) for prediction of a signal peptide and REP (Andrade et al., 2000)for the prediction of repeats. Additional annotation information was used from the AnoEST database and the Ensembl databases.

Identifier	Database	Notes
DNA & RNA		
ENSANGG00000010552	Ensembl	gene
ENSANGT00000013041	Ensembl	transcript
TCLAG004353	AnoEST	EST contig
protein		
ENSANGP00000013041	Ensembl	protein
EAA11514	Genbank	protein

Table 4.1. Gene, transcript and protein identifiers for *LRIM1*.

We used the Blastp programme to identify the best matching protein with a solved structure in the Protein Data Bank (PDB) and REP programme was used to predict the number of different LRR repeats (Andrade et al., 2000). The LRR repeats were aligned to the LRR repeats of the von Willebrand factor binding domain of glycoprotein Ib alpha using the ClustalW (Thompson et al., 1994) and the HMMER (Durbin, 1998) alignment programmes and was edited by hand to anchor the conserved key parts of the repeats. To account for 2 additional LRR repeats in the LRIM1 protein, a variant sequence of the model protein structure was made, by superimposing 2-4 repeats of the model structure to the 6-8 repeats of the same structure. Finally the aligned sequences were visualised with the ClustalW programme with a custom-made colouring scheme. The calculation of the three-dimensional model structure of the LRIM1 was performed with the MODELLER programme (Sali and Blundell, 1993) and visualised with RasMol programme. Secondary structures were coloured with the RASMOL default colouring scheme.

Mosquito rearing and species used

Details of the mosquito are provided in chapter 3. Three *A. gambiae* were used in this study: the G3 and Yaoundé strains, which are susceptible to *P.* berghei infection, and the L3-5 strain, refractory to *P. berghei* infections.

Generation of double stranded RNA for RNA interference

Fragments of LRIM1 and control genes LacZ and GFP were amplified by polymerase chain reaction (PCR) with oligonucleotide primers containing T7-promoter sequence overhangs (Table 4.2). The PCR products were checked for correct size in agarose gels, purified with the Qiagen PCR purification kit (Qiagen, Germany) and used as templates for the generation of dsRNA with the T7-Megascript kit (Ambion, USA). Reactions were carried out overnight (o/n) at 37° C, templates were DNAse treated and RNA products were Phenol-Chloroform purified, ethanol precipitated, resuspended in ultrapure H_2O, quantified and normalised to 3 µg/µl. DsRNA was checked on agarose gels prior to use. Mosquitoes that have been injected with dsRNA of *GFP* and *LRIM1* are referred to as *dsGFP* and *dsLRIM1* respectively, throughout the remainder of the thesis.

Primer name	Sequence
For generation of dsRNA	
GFP-T7-forward	5'-TAATACGACTCACTATAGGGCAAGACACGTGCTGAAGTCAA-3'
GFP-T7-reverse	5'-TAATACGACTCACTATAGGGGCCTGAATTTAACCAGGAACC-3'
LacZ-T7-forward	5'-TAATACGACTCACTATAGGGAGAATCCGACGGGTTGTTACT-3'
LacZ-T7-reverse	5'-TAATACGACTCACTATAGGGCACCACGCTCATCGATAATTT-3'
LRIM1-T7-forward	5'-TAATACGACTCACTATAGGGAATATCTATCTGCGAACAATAA-3'
LRIM1-T7-reverse	5'-TAATACGACTCACTATAGGGTGGCACGGTACACTCTTCC-3'
For RT-PCR	
LRIM1-1914F	5'-CATCCGCGATTGGGATATGT-3'
LRIM1-1983R	5'-CTTCTTGAGCCGTGCATTTTC-3'
S7-1	5'-GTGCGCGAGTTGGAGAAGA-3'
S7-2	5'-ATCGGTTTGGGCAGAATGC-3'

Table 4.2. The sequence of the primers used for generation of dsRNA and for RT-PCR. The dsRNA primers were engineered to produce PCR products of approximately 500 bp in size. T7 sequences are underlined.

RNA isolation and real-time PCR

Total RNA was isolated using the TRIzol Reagent (Invitrogen) according to the supplier's instructions, treated with DNAseI to remove and used to synthesize cDNA using the moloney murine leukemia virus reverse transcriptase and oligo $(dT)_{12\text{-}18}$

(Life Technologies, Inc.). Quantitative RT-PCR was performed with *LRIM1* specific primers (Table 4.2) using the SYBR Green PCR Master Mix kit (Applied Biosystems) according to the manufacturer's instructions and the ABI Prism 7000 Sequence Detection System. Relative gene expression values were calculated using the Comparative CT Method after checking for the efficiency of target amplification as described in the ABI Prism 7700 Sequence Detection System User Bulletin #2. The S7 ribosomal protein gene was used as internal reference (Table 4.2).

Injection of dsRNA to mosquitoes for RNAi assays

Functional analysis of gene knockdown experiments by RNAi was previously described (Blandin et al., 2002). Briefly, mosquitoes were anaesthetised by CO_2 administration and injected into the thorax with 69 nl of 3μg/μl of dsRNA with a Nanoject microinjector (Drummond Scientific, USA). After injection, mosquitoes were immediately returned to the insectary incubators (27° C, 75% relative humidity) and were allowed to recover. Dead mosquitoes were removed 24h after injection.

Bacterial infections of mosquitoes

For mosquito infections with Gram-negative (*Escherichia coli*) and Gram-positive (*Staphylococcus aureus*) bacteria, O/N bacterial cultures were used to inoculate subsequent cultures which were grown until reaching OD_{680nm} = 0.6. Bacteria were washed and diluted in phosphate buffered saline (PBS) to varying OD_{680nm} concentrations: 0.005, 0.01 and 0.05 for *E. coli* (1×, 2× and 10× respectively) and 0.004, 0.02, 0.04 and 0.08 for *S. aureus* (1×, 5×, 10× and 20× respectively). Prior to injection, bacterial samples were checked in the optical microscope to verify putative contamination by other bacteria.

Four days after dsRNA injection, mosquitoes were injected to the thorax with 69nl of fresh bacterial suspension of the tested concentration. Dead mosquitoes were removed the next day and not considered in the analysis, as it was doubtful whether their death was due to the bacterial infection or the severity of injection. The mosquito survival was scored every day for a period of 15 days by counting and removing dead mosquitoes. Data were used to construct survival curves with the non-parametric product limit estimator (Kaplan-Meier) method and the G^{ρ} test was used

to assess the statistical significance (P-values) of the difference of the Kaplan Meier curves. All calculations and graphical representations were performed in the R Statistical package (Team, 2006).

Infection of mosquitoes with Plasmodium parasites

For the laboratory infection of mosquitoes with the *P. berghei* parasite, the GFP-CON transgenic 259cl2 strain was used (Franke-Fayard et al., 2004), which was engineered to express the green fluorescent protein (GFP) protein through the entire parasite life cycle. The parasite was maintained in $CD1^+$ mice by blood passages. The parasitemia (parasite infected erythrocytes) of the mice was checked 3-6 days after passages under a light microscope by blood-smears counterstained with Giemsa solution. Mice with at least 10% parasitemia and visible gametocyte stages were used for infection experiments. Mice were anesthetised by intramuscular injection of a Xylazine: Ketamine: PBS (3:2:1 v/v) solution**.** Female mosquitoes were allowed to take an infectious blood meal on the anesthetised mice for 30-45 min at 19° C. Unfed mosquitoes were discarded 28h post infectious blood meal and fed mosquitoes were kept in 19° C to allow parasite development. All procedures involving animal work were carried out with special licenses from the respective institutions and under the strictest ethical criteria.

Determination of parasite load and statistical tests

To determine the number of fluorescent *P. berghei* parasites in infected midguts, mosquitoes were dissected 7-10 days after infectious blood meal and their midguts were fixed for 30-45 in 4% formaldehyde solution in PBS, washed 3 times with PBS for 20 min, mounted on microscope slides with Vectashield mounting medium (Vectorlabs, CA, USA) and observed with a UV-light fluorescent microscope (Zeiss, Heidelberg, DE).

Since oocyst numbers per mosquito midgut do not follow a normal distribution and do not have similar variances, appropriate non-parametric tests were used to determine the statistical significant difference of mean oocyst numbers. For pairwise comparisons between control and *LRIM1* dsRNA injected mosquitoes the Mann-Whitney (U-test) was used. For 3 or more comparisons, the Kruskal-Walis test

(analogous to a non-parametric ANOVA analysis) was performed and multiple comparisons (post hoc tests) were performed using the Mann-Whitney test. In all the calculations, results were considered statistically significant when displaying P-value less that 0.05. All calculations and the graphical plots were performed in the R statistical package (Team, 2006).

Generation of peptide antibodies against LRIM1 full-length protein

To predict candidate peptides for the generation of antisera that recognise the protein, the full length predicted protein sequence of LRIM1 was subjected to a variety of *in silico* tests to predict protein secondary structures and calculate their antigenic index. We used algorithms for the prediction of alpha helices, β-sheets, turns and coiled-coil domains (Chou-Fasman and Garnier-Robson algorithms), surface regions (Emini), amphipathic regions (Eisenberg), flexible regions (Karplus-Schulz), hydrophilic regions (Hopp-Woods and Kyte-Dollitle) and the antigenic index (Jameson-Wolf) that were implemented in the DNAstar software package (DNAStar, USA). Selected results of the algorithms for the full-length protein are presented in Fig. 4.7. Candidate peptides should be of relatively small size (14-16 aa) and have antigenic, hydrophilic and accessible properties. For the recognition of native protein, the peptides should be localised in the surface and strong secondary elements such as helices, sheets or turns should be avoided.

Peptide name	Peptide sequence	Location	BLASTP hits to proteome
garfield	EARSSKNAKRKMMS	5' of signal peptide	3 (1 insignificant hit)
buzz	EIKQNGNRYKIEKVTDS	5' of LRR domain	2 (1 insignificant hit)
oddie	HAANNNISRVSCSRGQGKKNIYLA	inside LRR structure	1
woody	KQTVKKLTGQ NEEEC	3' of LRR domain	4 (3 insignificant hits)
scroodge	RYEEMYVEQQSVQNNAIRDW	inside coiled-coil structure	23 (15 insignificant hits)

Table 4.3. The five candidate peptides for immunisations. Each peptide was assessed for homology to other known proteins of the *Anopheles* proteome.

The application of the above parameters identified 5 candidate peptide sequences (Fig. 4.7 and Table 4.3). Two additional criteria were used for the selection of appropriate peptides: i) whether the peptide was found inside known LRIM1 domains

and ii) whether the peptide sequence showed homology to other proteins in the *Anopheles* proteome (Table 4.3). The latter was assessed by alignment of the peptides to the proteome using the blastp programme with default parameters.

The 'Scroodge' peptide was eliminated due to its occurence inside the coiled coil domain of LRIM1 and the numerous significant hits to other *Anopheles* proteins. The 'Garfield' peptide was located 5' of the TM or signal peptide domain, making it an unsuitable candidate for fear of recognising the intracellular part in the event that the protein was transmembrane. 'Oddie' was more suitable as a candidate, as it displayed relatively high hydrophilicity and high specificity to the LRIM1 protein; however, its localisation inside the LRR domain could render the peptide inaccessible to the antibodies. Subsequently, 'Woody' and 'Buzz', two peptides flanking the LRR domain, were chosen for the generation of antibodies.

Peptides were synthesised chemically, coupled to m-Maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) and injected to rabbits, according to the supplier's protocol (Eurogentec, BE). After the third immunisation and terminal bleeding, antisera were affinity purified. Antisera were tested for their specificity and sensitivity to LRIM1 protein (in native and non-native conformations), as discussed in the results sections. To control for the antibody specificity, protein assays were performed using the preimmune serum (data not shown). Results from Western blot experiments with antisera raised against 'Woody' and 'Buzz' peptides were similar, and thus, 'Woody' was chosen for the experiments, unless otherwise stated.

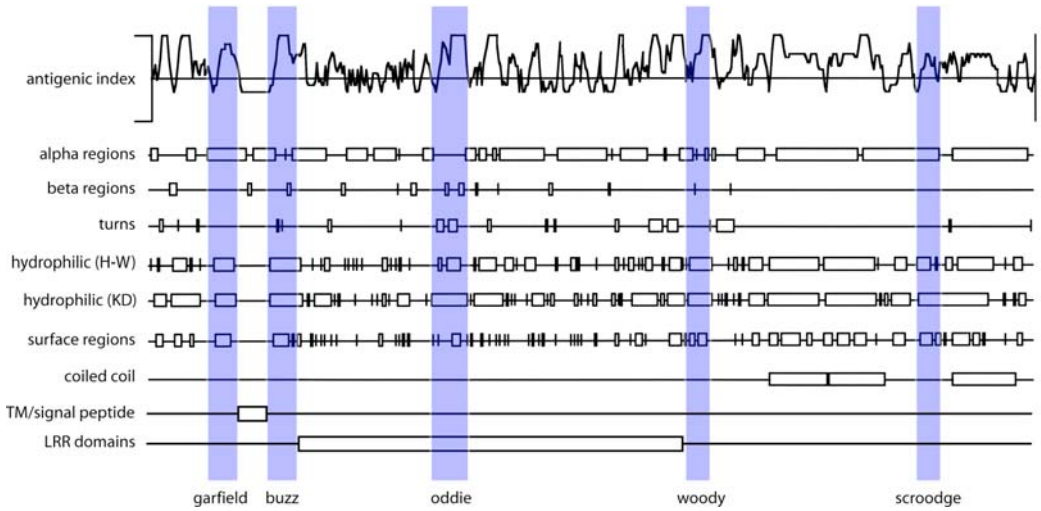


Figure 4.6. Secondary structure prediction domains for the selection of candidate antigenic peptides. Schematic representation of the antigenic index (graphical plot), secondary structures (boxed areas) predicted with selected algorithms and domain predictions (coiled coil, TM/signal peptide, LRR domain) of the full length LRIM1 protein. Blue boxes depict the positions of candidate peptides in relation to the domain structure of LRIM1.

Western blot analysis

For hemolymph collection, the proboscis of mosquitoes was cut with microdissection scissors and a small drop of hemolymph was collected in a pippete tip. The droplet was transferred immediately in protein loading buffer. For samples of midgut and carcass (the latter being the remainder of the mosquito after decapitation and midgut dissection), mosquitoes were dissected in ice cold PBS supplemented with a combination of protease inhibitors (Complete Protease Inhibitor Cocktail Tablets, Roche, CH). Tissues were homogenised and diluted in PLB solution (25 mM Hepes ph=7.5, 1.5 mM $MgCl_2$, 300 mM NaCl, 0.5 mM DTT, 0.2 mM EDTA, 0.1% Triton-X) supplemented with protease inhibitors. Western blot analysis was performed in 10% polyacrylamide gel and separated proteins were transferred to Hybond-P membranes (GE Healthcare, UK), blocked for at least 1h in 5% dry skimmed milk in PT (1× PBS, 0.1% Tween-20) and incubated O/N with 1:100 dilution of 'Woody' Ab in PT solution. Bound antibodies were detected with anti-mouse-conjugated horse-radish peroxidase Abs (in 1:10 000 dilution) by using a chemiluminescense kit (ECL, GE Healthcare, UK). Polyacrilamide gels were also analysed by Coomasie Brilliant Blue staining. All steps were performed as described (Sambrook and Russell, 2001).

Immunofluorescence

Midguts were dissected in ice-cold PBS supplemented with a combination of protease inhibitors (Complete Protease Inhibitor Cocktail Tablets, Roche, CH). Dissected midguts were prefixed in a fixative solution (4% formaldehyde, 1× PBS, 2mM $MgSO_4$, 1 mM EGTA, pH = 7.2) for 30-45 sec and transferred to ice-cold PBS for cutting open the midgut tissue. All subsequent steps were performed in room temperature (RT) unless indicated differently. The midguts were then placed in fixation solution for 30-45 minutes. After fixation, the blood bolus was carefully removed in PBS solution and midguts were further washed 3 times in PBS solution for 20 min and blocked in PBT solution (1× PBS, 1% BSA, 0,5% Triton-X) for at least 1h. Appropriate dilutions (Table 4.4) of primary Abs were incubated O/N in PBT solution at 4°C, washed three times with PBT for 20 min and incubated with appropriate of dilutions secondary antibodies in PBT for 4h (Table 4.4). After 3

washes with PBT for 20min, midguts were counterstained with $TOPRO_3$ for 5 min and mounted onto microscope slides with Vectashield mounting medium (VectorLabs, CA, USA).

Reagent name	Origin	Dilution	Details
woody antibody	custom peptide antibody, Eurogentec, BE	1:250	LRIM1 antibody
Alexa anti-Rabbit-568/ 643 conjugated antibodies	Invitrogen, CA, USA	1:1500	secondary antibodies
$TOPRO_3$	Invitogen, CA, USA	1:5000	nuclear staining
Phalloidin-568 conjugated	Invitogen, CA, USA	1:100	actin staining

Table 4.4. The antibody and reagent dilutions for the immunofluorescence experiments.

Visualisation was performed with a Leica SP2 confocal microscpe and image analysis using the LCS software (Leica Microsystems) and ImageJ programme (Abramoff et al., 2004). Additional image adjustments were performed with Adobe Photoshop CS2 software.

LRIM1 domain architecture and modelling of the LRR domain

The predicted protein sequence of LRIM1 was subjected to *in silico* prediction programmes to characterise its domain architecture. The results are summarised in Fig. 4.7. LRIM1 consists of a short N-terminal transmemembrane or signal peptide, an LRR domain corresponding to almost 40% of its residues and coiled-coil domains at its C-terminal. Both the LRR and coiled-coil domains are generic protein binding domains that have been identified in a variety of protein functional classes. It is thus likely for LRIM1 to mediate its function by binding to and interacting with other so far unknown proteins. In addition, LRIM1 could be implicated in pattern recognition, as it is known for other LRR-containing proteins, e.g. TLRs.

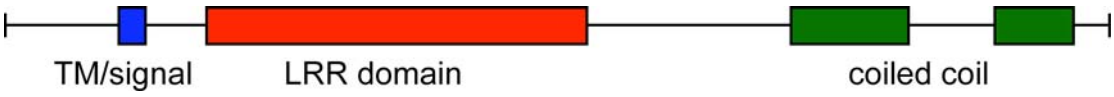


Figure 4.7. Schematic presentation of the LRIM1 protein domains. The sizes of the domains are drawn in scale to the full length protein sequence.

To determine the number of LRR repeats, the LRR domain sequence was analysed with the REP programme (Andrade et al., 2000) and the results are presented in Table 4.5. Six LRR repeats are predicted with high significance (LRR1-4, 6-7) and three repeats with lower significance (LRR5, 8,9). In addition, only a few LRR repeats (LRR2, LRR3, LRR7) show similarity with the typical LRR sequence, which codes for extracellular proteins. Two short peptides that do not correspond to *bona fide* LRR repeats were also identified. The first resides in the N-terminal part of the LRR domain and presumably corresponds to a degenerate repeat. However, several studies from other proteins have identified that sequence in the N-terminus of the LRR domain, which is contributing to the LRR binding specificity to the target protein (Bell et al., 2003; Huizinga et al., 2002; Kajava, 1998). Interestingly, the second short peptide is located in the middle of the LRR domain, is shorter that the typical LRR repeats and corresponds to the ‘oddie’ peptide, which was identified as a candidate for the generation of immune antisera for LRIM1 due to its high antigenicity and hydrophilicity (for details, refer to materials and methods).

Peptide name	Significance (P-value)	Sequence
N-terminal	-	EKVTDSSLKQALASLRQS
LRR1	8×10^{-7}	AWNVKELDLSGNPLSQISAADLAP
LRR2	4.92×10^{-5}	TKLELLNLSSNVLYETLDLESL
LRR3	6.5×10^{-5}	STLRTLDLNNNYVQELLVGPSIET
oddie	-	LHAANNNISRVSCSR
LRR4	1.5×10^{-4}	GQGKKNIYLANNKITMLRDLDEGCRS
LRR5	not significant	RVQYLDLKLNEIDTVNFAELAAS
LRR6	7×10^{-5}	SDTLEHLNLQYNFIYDVKGQVVF
LRR7	6.93×10^{-7}	AKLKTLDLSSNKLAFMGPEFQS
LRR8	not significant	AGVTWISLRNNKLVLIEKALRF
LRR9	not significant	LEHFDLRGNGFHCGTLRDFFSK

Table 4.5. The proposed LRR repeat. Sequence predictions by the REP programme are obtained by homology based methods and estimates of statistical significance (P-values). A provisional name has been given to the peptides of the LRR domain. The two peptides with missing P-values do not correspond to *bona fide* LRR repeats

The three dimensional structure of the LRR domain of LRIM1 was modelled on a similar LRR domain of the von Willebrand factor binding domain of the glycoprotein Ib-β domain (PDB: 1m0Z). To account for additional LRR repeats in LRIM1 than the model sequence, a modified sequence of the von Willebrand factor binding domain was produced with two additional LRR repeats. The leucine residues between those two sequences were aligned together, although little or no conservation was observed between the remaining residues (Fig. 4.8A). Accordingly, arrangement of the LRR domain in the three-dimensional space (Fig. 4.8C) produced a horse-shoe shaped structure that is characteristic of the von Willebrand factor binding domain of glycoprotein Ib-β domain and other LRR containing proteins. However, in both structures we observed a deviation from the proposed LRR architecture: while β-sheets are lining the inner circumference of the horse-shoe structure (Fig. 4.8B,C, yellow colour), very few a-helices are observed in the outer circumference (Fig. 4.8B,C, magenta colour). Interestingly, although 'oddie' is not a *bona fide* LRR peptide, the model suggests that it could potentially be a repeat, as it fits the model without many distortions (Fig. 4.8C, white arrowheads). Conversely, if this peptide is excluded from the LRR model, it would form an insertion that would divide the LRR peptide in two parts and could influence putative LRR binding to target proteins. Such a phenomenon has been described for other LRR-containing proteins and could possibly explain the higher antigenicity and hydrophilicity of this peptide compared to the remaining LRR structure (Fig. 4.6). However, in the absence of any other

structural information, whether ‘oddie’ contributes to the LRR sequence or creates an insertion dividing it into two LRR domains remains to be elucidated.

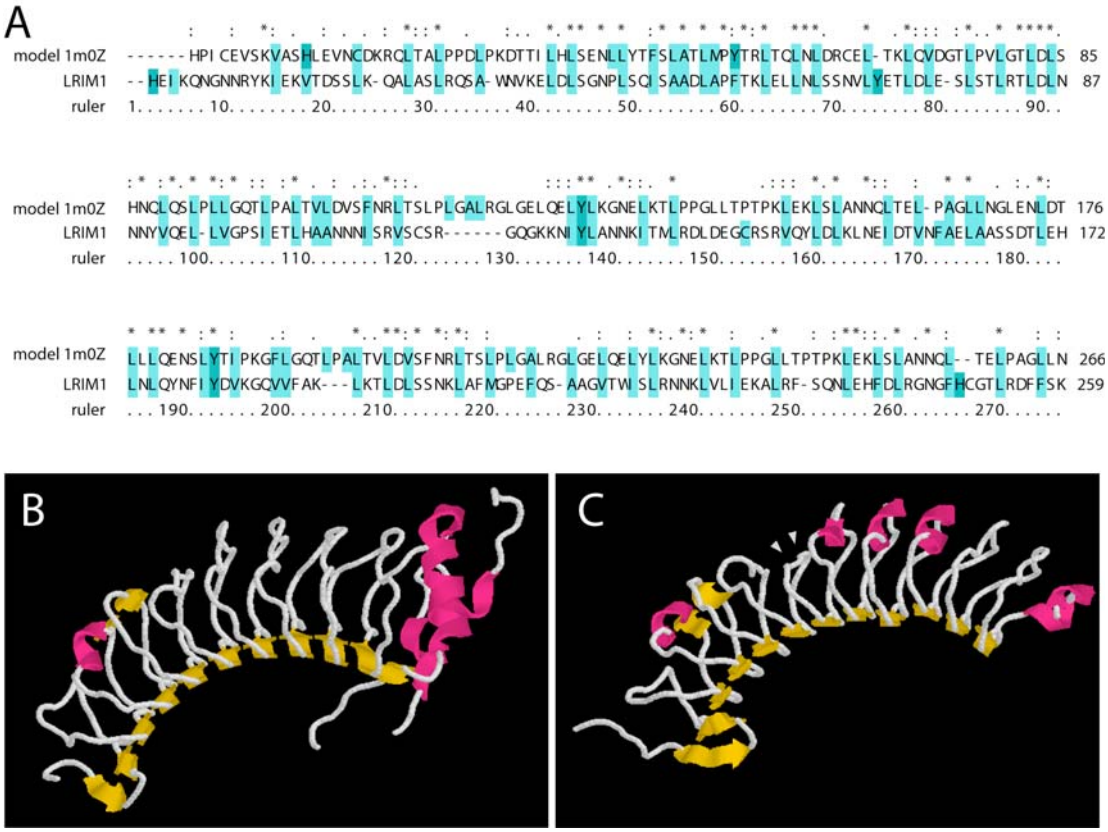


Figure 4.8. Modelling of the LRR domain of LRIM1. A) Alignment of the LRIM1 LRR domain sequence to the modified sequence of the von Willebrand factor-binding domain of Glycoprotein Ib alpha. Coloured residues show the conservation of hydrophobic aminoacids (L, V, A) and the anchorage of the leucine residues between the two sequences (alignment is visualised with Clustalx). Three-dimensional structures of the LRR domains of the (B) von Willebrand factor of glycoprotein Ib alpha (PDB:1m0Z) and (C) LRIM1. Models were plotted with the RasMol programme and secondary structures (a-helices in magenta and b-sheets in yellow) are shown. Arrowheads display a shorter structure corresponding to the oddie peptide.

LRIM1 temporal and spatial expression

The global gene expression profiles during the *Anopheles* life cycle and in adult mosquito tissues (which are presented in detail in chapter 3) were used to determine the temporal expression profile of *LRIM1*. Results from averaging the expression profiles of 6 ESTs corresponding to *LRIM1* in 4 biological replicates are presented in Fig. 4.9A,B. The expression profiles of individual ESTs are presented in supplementary material (Fig. 4.S1, S2). *LRIM1* expression is at its lowest level in the embryonic time period, increases at the successive larval time periods and peaks at the Ld period. This expression pattern is reminiscent of that of proteloytic enzymes,

as has been discussed in chapter 3 (Fig. 3.9C1). However, expression of *LRIM1* differs from this pattern in that it declines in the Le and the P time period and then increases again at the adult male and female time periods. *LRIM1* has been included in the SOM cluster 15 and the embryo-low pupae low developmental pattern, EOpm (Fig. 3.8).

The expression pattern of *LRIM1* during the mosquito lifecycle was compared to expression profiles of other genes. Using two different correlation coefficients, we detected 14 and 11 TCLAG contigs that share more than 0.9 similarity with *LRIM1* (refer to Table 4.6 and supplementary Fig. 4.S2 for the expression profile each contig relative to expression of *LRIM1*). Interestingly, some TCLAG contigs in those lists correspond to ENSEMBL gene models with putative innate immune functions. Among them, *LRIM2* and *CLIPA2* and ENSANGG00000017669 are shared in both lists, indicating a great degree of similarity with *LRIM1*. *LRIM2*[4] is also upregulated after *Plasmodium* infection and was recently shown to have a role in innate immune reactions against the *P. berghei* parasite (Riehle et al., 2006). *CLIPA2*, a member of the CLIPA domain of serine proteases, which share a CLIP domain but do not have a functional serine protease domain; however CLIPA2 KD does not result an an increase in parasite numbers (Michael Osta, personal communication). Interestingly, ENSANGG00000017669 is also located between CLIPA1 and CLIPA7, but was omitted by the recent gene prediction of Ensembl. The list also includes two other putative proteases (ENSANGG00000018929, ENSANGG00000020433) and genes involved in oxidation (ENSANGG00000012499, ENSANGG00000015133). TEP5 is a member of the family of complement-like thioester containing proteins, a member of which, TEP1, plays an important role in bacterial and parasite immunity (Blandin et al., 2004). Studies have previously shown that genes with similar expression pattern also share functional relevance (Eisen et al., 1998; Ge et al., 2001; Jansen et al., 2002; Lee et al., 2004). Therefore, the striking similarity of *LRIM1* expression with many members of the CLIPA proteases and other putative proteases merits further investigation.

Furthermore, expression profiling in mosquito adult tissues has shown that *LRIM1* is abundantly expressed in the head and the carcass but displays low expression in the midgut and the ovaries. *LRIM1* has been included in the list of genes belonging to the

[4] ENSANGG00000012041 has two gene name synonyms: LRIM2 and APL1.

5th co-expression cluster of adult tissues (Fig. 3.10), which display a characteristic dual high expression in head and carcasses. These profiles in adult tissues is consistent with those reported in another study (Marinotti et al., 2006), which interrogated the genome expression in adult tissues in female mosquitoes 24h post blood meal utilising a different microarray platform. Another microarray transcriptomic study using the Affymetrix GeneChip platform showed increased expression of *LRIM1* in mosquito hemocytes (S.B. Pinto, A. Koutsos, K. Michel and F.C. Kafatos, unpublished results). In this study, *LRIM1* belongs to a co-expression cluster, in which many other innate immune molecules (e. g. CLIPA,B and C serine proteases, C-type lectins and thioester containing proteins) have been identified.

In addition to LRIM1 expression in the developmental time periods and adult tissues, publicly available microarray information shows *LRIM1* transcript abundance at various time periods post blood meal (Fig. 4.9C). *LRIM1* is transiently increased 3h after the blood meal, gradually declines in a period of 1d to 3d and increases again 4d after blood meal. Interestingly, this high expression is also observed 15d after the blood meal.

Similarity (coefficient)	TCLAG contig (AnoEST v. 5)	ENSEMBL gene (Common gene name)
Pearson correlation		
0.98	TCLAG035575	ENSANGG00000017669
0.978	TCLAG033173	ENSANGG00000014360
0.965	TCLAG033134	ENSANGG00000018727 (TEP4)
0.964	TCLAG005467	ENSANGG00000012041 (LRIM2)
0.949	TCLAG032832	ENSANGG00000020433
0.946	TCLAG035574	ENSANGG00000017763 (CLIPA2)
0.944	TCLAG030516	ENSANGG00000017773 (CLIPA1)
0.942	TCLAG005478	ENSANGG00000018929
0.939	TCLAG010314	ENSANGG00000004547
0.924	TCLAG025458	ENSANGG00000015133
0.915	TCLAG012195	-
0.913	TCLAG005606	-
0.908	TCLAG049743	ENSANGG00000020522
0.902	TCLAG024123	-
Spearman correlation		
0.983	TCLAG005467	ENSANGG00000012041 (LRIM2)
0.967	TCLAG035574	ENSANGG00000017763 (CLIPA2)
0.967	TCLAG035575	ENSANGG00000017669
0.967	TCLAG024123	-
0.95	TCLAG025918	-
0.917	TCLAG030514	ENSANGG00000017677 (CLIPA6)
0.917	TCLAG011527	ENSANGG00000012499
0.917	TCLAG028986	ENSANGG00000008759
0.917	TCLAG039854	ENSANGG00000008759
0.917	TCLAG019968	-
0.917	TCLAG005606	-

Table 4.6. TCLAG contigs showing similar expression to LRIM1 in the mosquito life cycle. For the comparison, Pearson and Spearman correlation coefficients have been used and only TCLAG contigs showing more that 0.9 similarity are shown. For each TCLAG contig, the overlapping ENSEMBL gene model and the common name (whenever available) is noted.

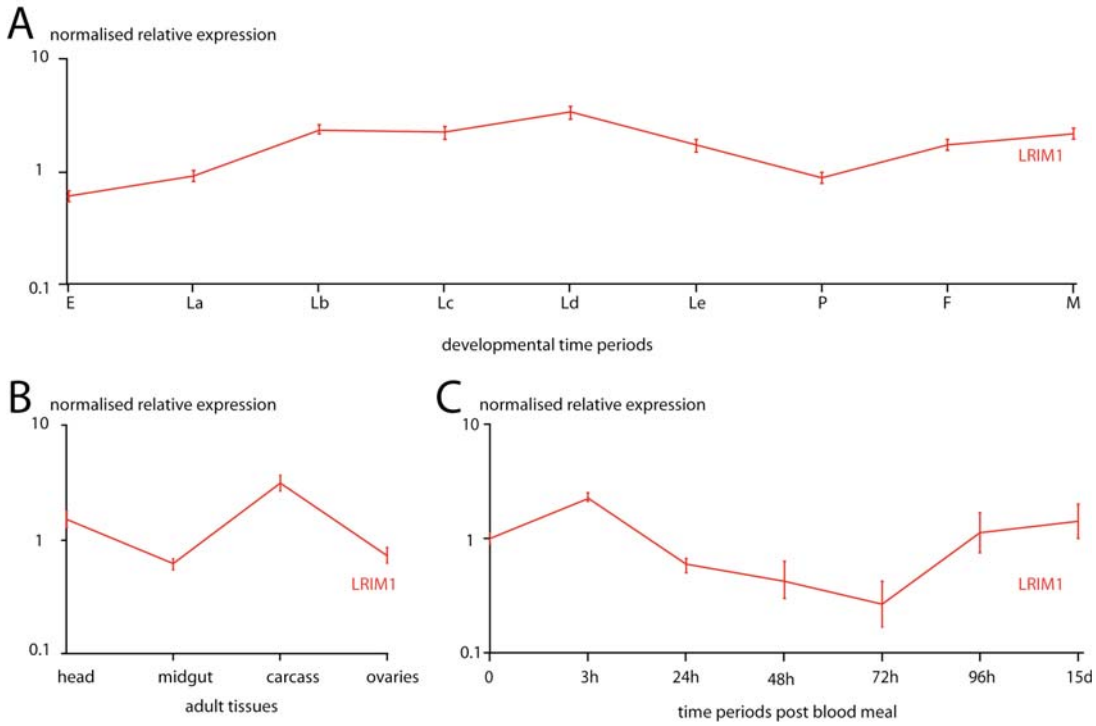


Figure 4.9. LRIM1 RNA temporal and spatial expression. Micrroarray expression profiles of in A) developmental time periods, B) adult tissues (both obtained by MMC1 microarrays) and C) blood meal periods (adapted from (Marinotti et al., 2005)). Normalised log10-transformed expression values are shown in y axes and vertical bars show standard error of expression in the respective samples.

Microarray analyses in mosquitoes indicated a rapid upregulation of *LRIM1* 24h after *P. berghei* infections ((Dimopoulos et al., 2002), Fig. 4.3). *LRIM1* expression after parasite infection was assessed by real time PCR (RT-PCR) in bloodfed and infected mosquitoes (Fig. 4.10) (Osta et al., 2004). Compared to blood feeding, *LRIM1* is induced in midguts and to a greater extent in carcasses 24h after parasite infection (Fig. 4.10A), a period which coincides with the majority of ookinetes invading the mosquito midgut epithelium. A more detailed analysis of the expression of *LRIM1* in mosquito midguts reveals that this transient upregulation is specific to the period of 24 and 28h post infection; significantly lower amounts of LRIM1 are detected at time periods before (18-20) and after (32-32) parasite infection; no difference is observed between bloodfed and infected mosquitoes (Fig. 4.10B).

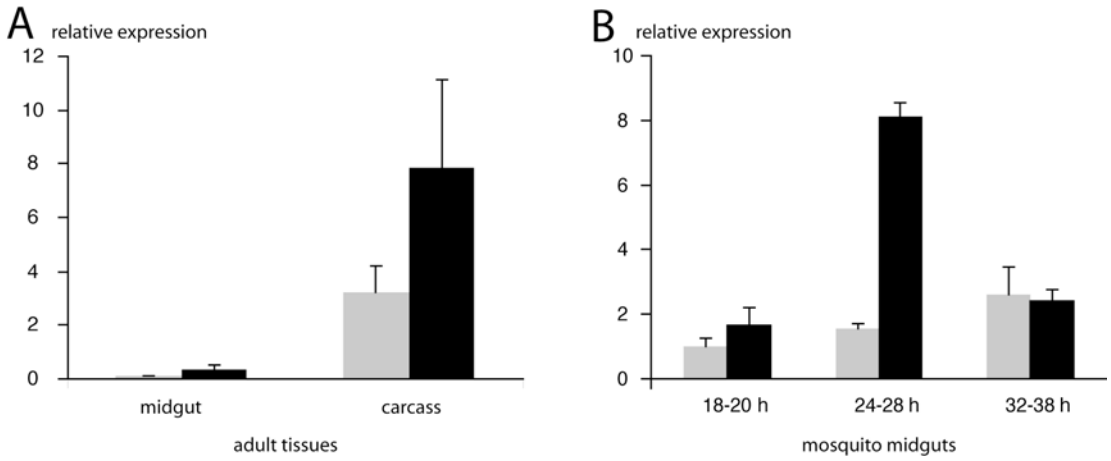


Figure 4.10. *LRIM1* RNA expression in midguts and carcasses. RT-PCR results of A) adult tissues and B) various time periods in mosquito midguts. Grey bars indicate blood meal and black bars indicate infection with *P. berghei* parasite. Data adapted from (Osta et al., 2004)).

For LRIM1 protein expression, rabbit polyclonal antibodies against a specific peptide sequence of LRIM1 (woody peptide, see material and methods for more information) were produced and used to monitor protein expression in tissue samples before and after parasite infection. A single band corresponding to a protein of approximately 50kD in size, which is in agreement with the expected protein size, was detected in the mosquito hemolymph. As mentioned earlier, data from microarray analyses indicated a hemocyte-specific origin for LRIM1. Results from the protein studies suggest that *LRIM1* is expressed in mosquito hemocytes and secreted into the hemolymph. The rapid upregulation of *LRIM1* RNA in mosquito midguts and carcasses could be explained by the fact that during *P. berghei* invasion hemocytes are known to be recruited to the mosquito midgut and are thus co-isolated with these tissues during sample preparations. However, no LRIM1 protein was detected in Western-blot assays of midgut tissues or carcasses (Fig. 4.11B), indicating that there might be either a rapid turnover of the protein due to the effect of midgut proteases, which are not inhibited by the protein inhibitors present during tissue isolation, or that the protein concentration in sessile haemocytes (before its secretion into the hemolymph) is significantly lower than the Western blot detection capability.

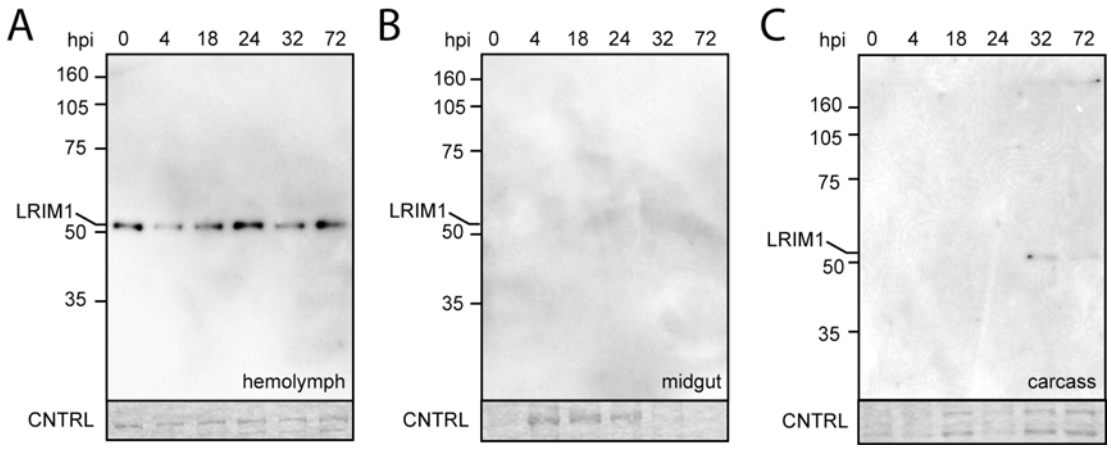


Figure 4.11. LRIM1 protein expression in mosquito tissues. Western blots of hemolymph (A), midgut (B) and carcass (C) protein extracts of mosquitoes at various hours after P. *berghei* infection. LRIM1 was detected above 50kD, which is in agreement with its expected size. The faint bands depicted at 32 and 72 hpi in carcasses are likely due to hybridisation atefacts. Total protein extracts (selected regions in coomasie staining) are used as a loading control (CNTRL).

Effects of LRIM1 *on mosquito survival after bacterial infection*

Two studies have provided a link between *LRIM1* and innate immune responses directed against bacteria. In the first study, knockdown of *LRIM1* by double stranded RNA (dsRNA) resulted in reduced phagocytosis of the Gram-negative bacterium *E. coli* in adults mosquitoes (Moita et al., 2005). In the second study knockdown of *REL2* -an NF-κB transcription factor orthologous to the Drosophila *Relish*- in cell lines resulted in downregulation of *LRIM1* (Meister et al., 2005).

To determine the possible role of *LRIM1* in bacterial immunity, variable concentrations of Gram-negative (*E. coli*) and Gram-positive (*S. aureus*) bacteria were injected into the hemolymph of either dsGFP or dsLRIM1 injected mosquitoes and their survival was scored for a period of 15 consecutive days. Survival curves according to the Kaplan-Meier statistical method were plotted for pairwise comparisons of control versus *dsLRIM1* mosquitoes in each tested concentration of *E. coli* (Fig. 4.12) and *S. aureus* (Fig. 4.13).

While the survival probability of the *LRIM1* KD mosquitoes was not significantly compromised using 1× concentration of the Gram-negative bacterium *E. coli* (Fig. 4.12A), increasing concentrations of this bacterium resulted in reduced mosquito survival in the *dsLRIM1* mosquitoes. In particular, the 1× concentration was lethal only after 7 days, as compared to the survival of the GFP dsRNA treated mosquitoes, whereas increasing concentrations resulted in almost immediate (1-3 days) killing of mosquitoes, leading to 63% and 54%.decrease in survival at day 16 of the experiment

for the 2× and 10× concentrations, respectively (Fig. 4.12B,C). The difference of the *LRIM1* KD survival curves in different bacteria concentrations was highly significant ($P < 0.001$) suggesting a concentration-dependent effect of *LRIM1* in mosquito resistance to *E. coli* infections (Fig. 4.S3B). In contrast, the survival curves of *GFP* dsRNA treated mosquitoes in the different bacterial concentrations did not show any statistical significance (Fig. 4.S3A).

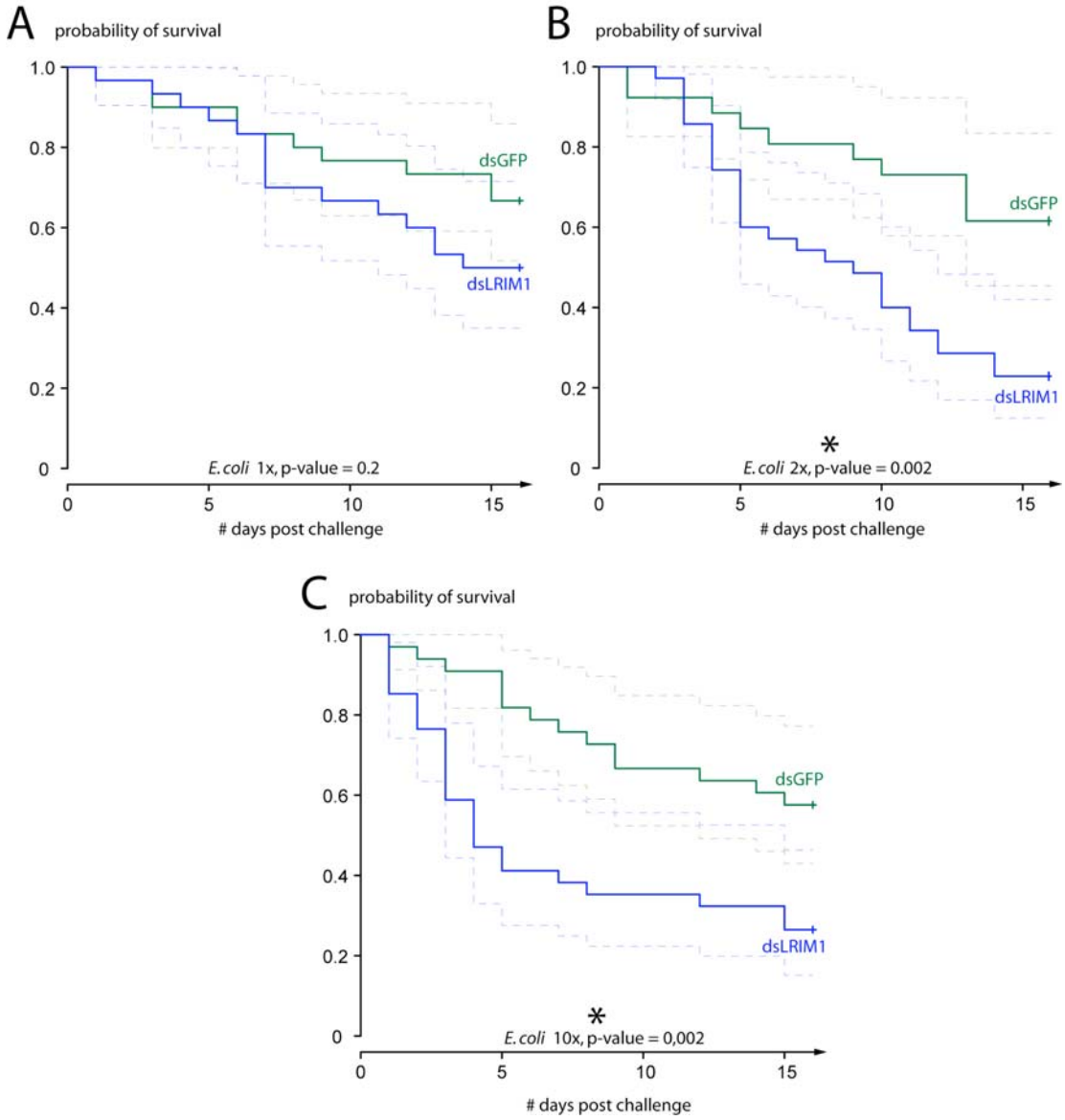


Figure 4.12. Survival of *dsLRIM1* knockdown mosquitoes after *E. coli* challenge. Kaplan- Meier curves of control and *dsLRIM1* mosquitoes after infection with 1× (A), 2× (B) and 10× (C) concentrations of *E. coli*. Solid lines represent the probability of survival, dotted lines represent the 95% confidence intervals and asterisks represent statistical significant difference of comparison of *dsLRIM1* knockdown against dsGFP mosquitoes.

A different response was observed after infection with the Gram-positive *S. aureus* bacterium as mosquito survival was drastically reduced even at low concentrations (1×) (Fig. 4.13A) and increasing concentrations did not have any further impact (5× and 10×, Fig. 4.13B,C respectively). In all concentrations, survival probability was greatly reduced for the first day of the experiments and the decrease in survival probability at day 16th was 52%, 63% and 88% for each bacterium concentration, respectively. In greater concentrations (20×), the survival of both control and dsLRIM1 mosquitoes was severely compromised and no statistical difference was detected, indicating that the immune system is saturated with *S. aureus* bacteria, such that it is unable to provide effective protection. When the four different concentrations of *S. aureus* are compared in dsGFP mosquitoes (Fig 4.S4A), a significant difference was observed, which was mostly ascribed to the 20× concentration. In contrast, the curves corresponding to the four different concentrations in the *dsLRIM1* mosquitoes are mostly similar (Fig. 4.S4B). Thus, in the *dsLRIM1* mosquitoes, survival probability is greatly compromised in the presence of *S. aureus* and this effect is concentration-independent. Nevertheless, in greater concentrations (20×), the high bacterial load causes death, irrespective of the absence of *LRIM1*.

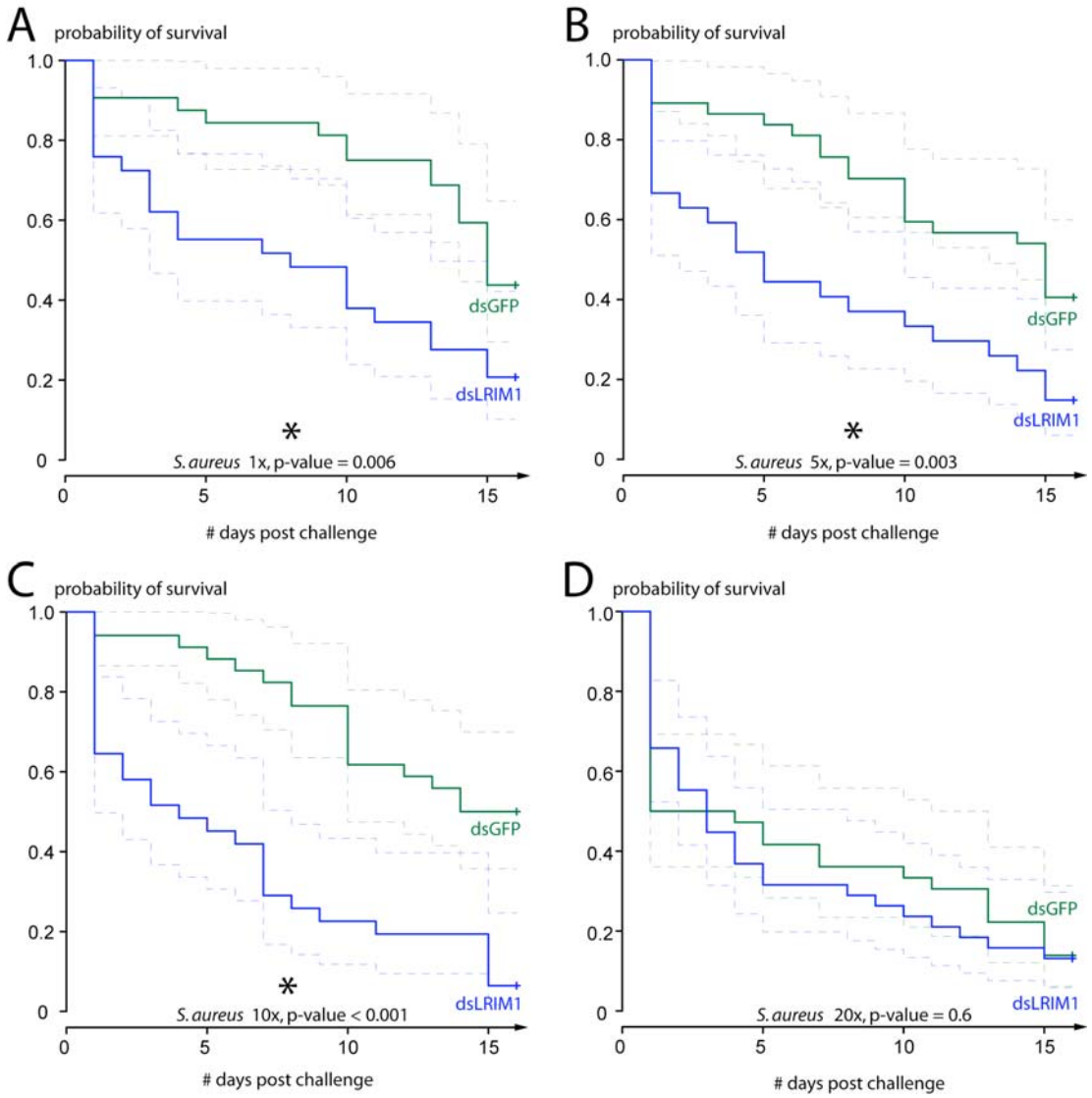


Figure 4.13. Survival of dsLRIM1 knockdown mosquitoes after *S. aureus* challenge. Kaplan- Meier curves of control and dsLRIM1 mosquitoes after infection with 1× (A), 5× (B), 10× (C) and 20× (D) concentrations of *E. coli*. Solid lines represent the probability of survival, dotted lines represent the 95 % confidence intervals and asterisks represent statistical significant difference of comparisons of LRIM1 knockdown against dsGFP mosquitoes.

Effects of LRIM1 on Plasmodium parasite development

As mentioned previously, *LRIM1* expression is strongly induced in mosquito midguts 24h after parasite infection with the rodent malaria parasite, *P. berghei*, a period that coincides with the majority of the ookinetes invading the epithelium, indicating *LRIM1* involvement in parasite infection responses (Dimopoulos et al., 2002; Vlachou et al., 2005). A previous study (Osta et al., 2004) in the susceptible (S) mosquito strain, observed a dramatic increase in the number of developing oocysts in

the dsLRIM1 knockdown mosquitoes compared to the control ones seven days after infection (Fig 4.14A,B). In the refractory (R) strain, the *P. berghei* parasites are known to rapidly melanise and remain in the midgut wall of the mosquitoes (Fig 4.14C). Melanisation is a well known insect immune reaction which in this strain causes a complete malaria transmission blockade. Strikingly, no melanisation was observed after LRIM1 KD and all the parasites develop into fluorescent oocysts (Fig. 4.14D). These oocysts are able to produce sporozoites that invade the salivary glands and can be transmitted into rodents after a mosquito bite (data not shown). Thus, absence of *LRIM1* is sufficient to revert the refractory phenotype to a phenotype similar to the susceptible mosquitoes. Furthermore, a dramatic increase in parasite numbers in the *dsLRIM1* knockdown mosquitoes compared to the control ones was observed (Fig 4.14C, D). Taken together these results suggest that *LRIM1* is critical for parasite killing in both S and R mosquitoes and for parasite melanisation in the R mosquitoes.

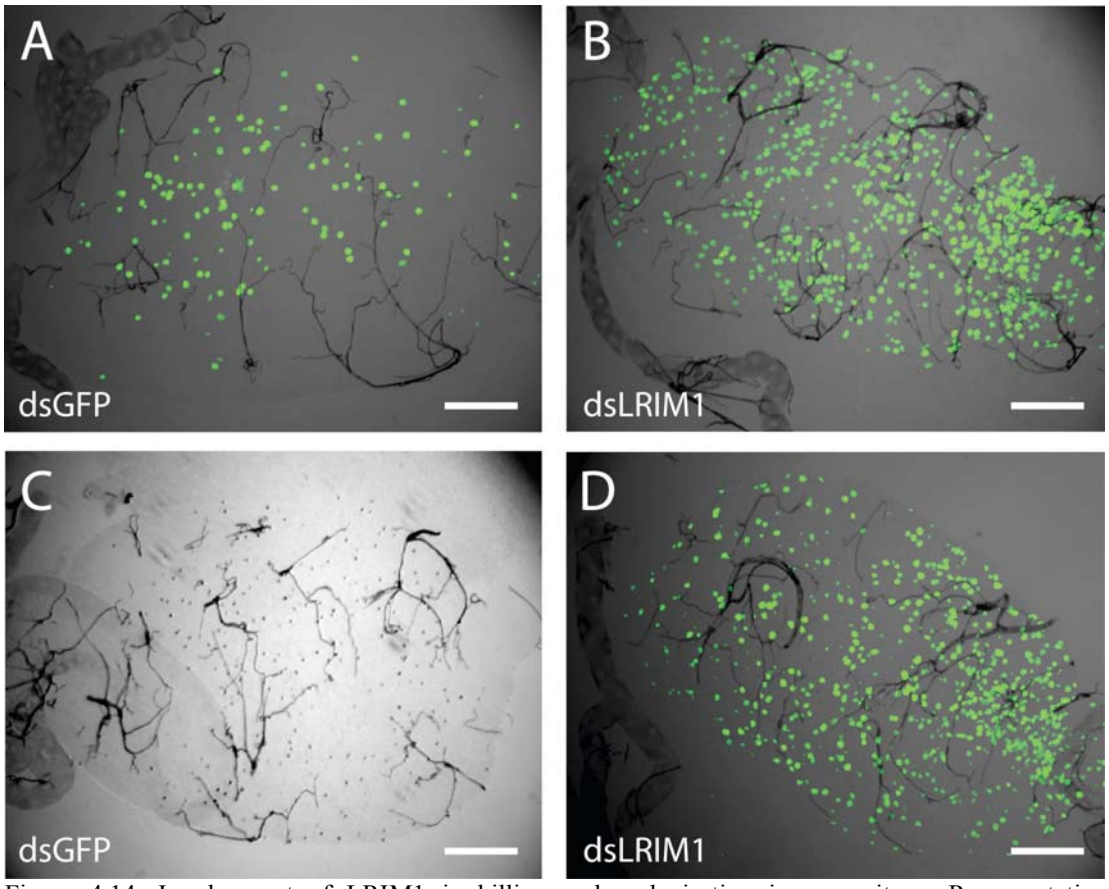


Figure 4.14. Involvement of LRIM1 in killing and melanisation in mosquitoes. Representative microscopic pictures of midguts of susceptible (A, B) and refractory (C, D) strains of mosquitoes. Green dots represent GFP fluoresecent parasites and black dots represent melanised parasites. In the absence of LRIM1, there is a great increase in the number of parasites developing in the midguts in the susceptible strain, and there is reversal of melanisation phenotype and an increase in the parasite numbers in the refractory strain. (scalebar = 50 μm).

We quantified the difference in parasite numbers between control and LRIM1 KD S and R mosquitoes by counting the number of oocysts and melanised ookinetes in mosquito midguts at 7-9 days post infection. Overall a highly significant (P-values < 0.001 for both the S and R mosquitoes) four-fold increase of the oocyst numbers was observed (Fig. 4.15A,B and Table 4.7), although results from individual experiments showed a varying increase (Fig. 4.S5 for the S mosquitoes and Fig. 4.S6 for the R mosquitoes), In addition, a substantial fraction of *LRIM1* KD midguts carried large numbers of oocysts (>300 oocysts; compare oocyst range between dsGFP and dsLRIM1 mosquitoes in Fig. 4.15A,B and Table 4.7). In R mosquitoes we additionally observed a slight increase in parasite prevalence (numbers of infected midguts, Table 4.7).

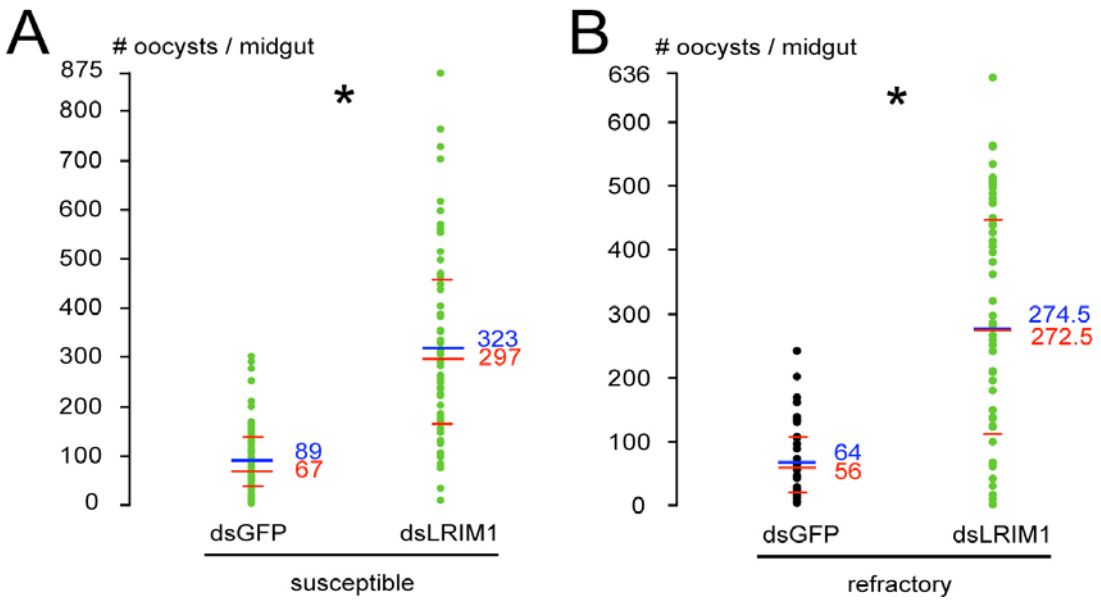


Figure 4.15. Effects of LRIM1 KD in parasite numbers in mosquito midguts. Comparable increase in the developing parasites is observed in the refractory strain (A) and the susceptible strain (B). Green dots show fluorescent (living) parasites, black dots melanised (killed) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisks denote statistically significant difference between control and dsLRIM1 mosquitoes (Mann-Whitney test, P-value <0.05). The results for the susceptible mosquitoes have been reanalysed from (Osta et al., 2004).

The similar phenotypes of the *LRIM1* KD mosquitoes between strains prompted us to directly compare results in oocyst numbers between different strains. Therefore, the KD experiments were repeated by feeding both R and S mosquitoes to the same *P. berghei* infected rodent, thus exposing both strains to the same parasite source and load.

Name	# exp	# mid	Preval.	Mean oocysts/ midgut (±SE)	Range	Fold difference	P -value
Susceptible strain (S)							
dsGFP	5	63	100%	89.87 (9.15)	3 - 301		$6.54\ 10^{-14}$
dsLRIM1		60	100%	323.3 (25.34)	7 - 875	3.6	
Refractory strain (R)							
dsGFP	3	38	84.21%	64.81 (9.84)	0 - 237		$3.98\ 10^{-7}$
dsLRIM1		52	90.38%	274.86 (26.37)	0 - 636	4.24	
Comparison of strains							
Susc. dsGFP	6	99	87.87%	90.57 (9.59)	0 - 400		$1,272\ 10^{-10}$
Susc. dsLRIM1		86	94.18%	202.03 (19.68)	0 - 773	2.23	
Ref. dsGFP		72	68.05%	58.37 (10.33)	0 - 555		
Ref. dsLRIM1		59	89.83%	167.5 (18.71)	0 - 530	2.87	

Table 4.7. Cumulative results of LRIM1 KD in S and R mosquito strains. In each experiment, information includes number of experiments, total number of mosquito midguts, prevalence of infection (% of infected midguts), mean oocyst numbers per mosquito midgut and standard error and range of oocysts in midguts (min – max). Fold difference is calculated for each pairwise comparison between LRIM1 KD and control (GFP dsRNA trated) mosquitoes. P-values correspond to the Mann-Whitney tests for pairwise comparisons and the Kruskal-Wallis test for comparison of all four types of mosquitoes.

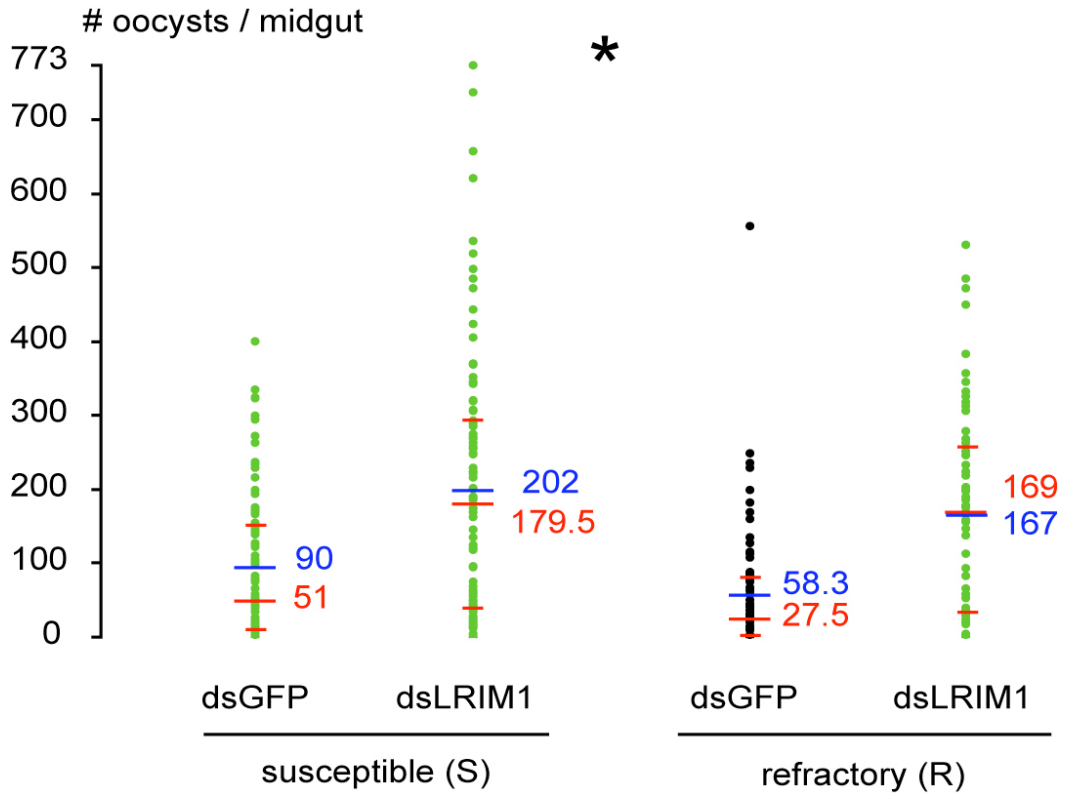


Figure 4.16. Comparison of oocyst numbers in susceptible and refractory mosquitoes infected with the same parasite numbers. Green dots show fluorescent (living) parasites, black dots melanised (killed) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisk denotes statistically significant difference between strains and control and dsLRIM1 mosquitoes (Kruskal-Wallis test, P-value < 0.001).

Similarly, there was an overall – albeit slight lower that our previous experiments – fold increase in parasite numbers in the ds*LRIM1* mosqutoes in both strains relative to the *dsGFP* (Fig. 4.16). A substantial increase in infection prevalence was also observed. However, by examining the individual experiments, we deduced that the overall difference is due to an observed increase in specific experiments: in the R mosquitoes a significant difference is detected in only two experiments (Table 4.8, experiment 2 and 6) whereas in the S mosquitoes only in tthree experiments (Table 4.8, experiments 1,3,4). Interestingly, experiment 5 (Table 4.8) did not show any statistical significant fold change in both mosquito strains. These results suggest that there might be other factors e.g. the intensity of parasite infection, that might influence the fold change in parasite numbers in these experiments. When the same data are anaylysed as pairwise comparisons between either the *dsLRIM1* or the *dsGFP* mosquitoes (Table 4.9), results show that comparable parasite numbers are observed in the *dsLRIM1* treated mosquitoes in the two strains but not in the control mosquitoes. Possibly, the immune response that entails *LRIM1*-mediated killing is activated in different parasite concentrations and in different time periods between the two strains, explaining both the variability of fold changes between experiments and experiments where the difference between *dsLRIM1* and control mosquitoes is not significant. In the absence of an immune response that entails *LRIM1*, when the mosquito strains are fed with the same parasite numbers no phenotypic and quantitative differences are detected, indicating that *LRIM1* may, at least partly, account for such differences.

Exp.	Mean # oocyst/ midgut		Fold change	P-value	Mean # oocyst/midgut		Fold change	P-value
	Susceptible dsGFP	Susceptible dsLRIM1			Refractory dsGFP	Refractory dsLRIM1		
1	122.43	217.85	1.78	(0.094)	49.33	299.11	6.06	*0.0016
2	8.00	34.64	4.33	*0.002	4.67	12.00	2.57	(0.4363)
3	144.56	135.31	0.94	(0.666)	41.40	118.50	2.86	*0.0306
4	130.30	211.53	1.62	(0.891)	28.80	206.64	7.17	*0.0101
5	68.64	418.57	6.10	(0.417)	178.50	317.80	1.78	(0.7446)
6	34.73	176.21	5.07	*0.047	80.04	149.93	1.87	(0.0716)

Table 4.8. Pairwise comparisons of the oocyst numbers between the control and dsLRIM1 mosquitoes in the susceptible and refractory mosquitoes. For each experiment, the average number of oocysts/midgut, the fold difference between dsLRIM1 and dsGFP mosquitoes and the P-values of the Mann-Whitney tests are noted. Asterisks denote statistical significant difference(P-value < 0.05) and parentheses denote not significant difference (P-value ≥ 0.05).

Exp.	mean # oocyst/ midgut		Fold change	P-value	mean # oocyst/midgut		Fold change	P-value
	Susceptible dsGFP	Refractory dsGFP			Susceptible dsLRIM1	Refractory dsLRIM1		
1	122.43	49.33	0.40	*0.03	217.85	299.11	1.37	(0.29)
2	8.00	4.67	0.58	(0.32)	34.64	12.00	0.35	*0.01
3	144.56	41.40	0.29	*0.002	135.31	118.50	0.88	(0.93)
4	130.30	28.80	0.22	*0.01	211.53	206.64	0.98	(0.91)
5	68.64	178.50	2.60	(0.36)	418.57	317.80	0.76	(0.34)
6	34.73	80.04	2.30	*0.05	176.21	149.93	0.85	(0.91)

Table 4.9. Pairwise comparisons of the different mosquito strains in the control and dsLRIM1 mosquitoes. For each experiment, the average number of oocysts/midgut, the fold difference between the strains and the P-value of the Mann-Whitney tests is noted. Asterisks denote statistical significant difference (P-value < 0.05) and parentheses denote not significant difference (P-value ≥ 0.05).

In addition to the melanising phenotype in the R strain of *A. gambiae*, studies have shown that knockdown of two members of the C-type lectin family, CTL4 and CTLMA2, produces a similar melanisation phenotype in the S strain of mosquitoes (Osta et al., 2004). Ongoing work aims to establish whether the mechanism of melanisation in those two strains is similar. To determine whether absence of LRIM1 is sufficient to revert the phenotype of CTL4 or CTLMA2 KD mediated melanisation in the S strain, double knockdown experiments were performed (Osta et al., 2004). Recently, a stdy determined that the CTL4-induced melanisation in the S mosquitoes directly kills ookinetes, whereas melanisation in the R mosquitoes merely disposes dead parasites (Volz et al., 2006). The results indicated that *LRIM1* is epistatic to *CTL4* and *CTLMA2*; in the double KD experiment, absence of *LRIM1* abolishes melanisation and the mosquitoes show similar number of fluorescent oocysts as the *dsLRIM1* KD mosquitoes. Taken together, the results suggest that *LRIM1* is involved in two important innate immune responses against the parasite: in killing and melanisation, the latter both in the observed spontaneous melanisation of the R strain or the C-type lectin KD-induced melanisation of the S strain.

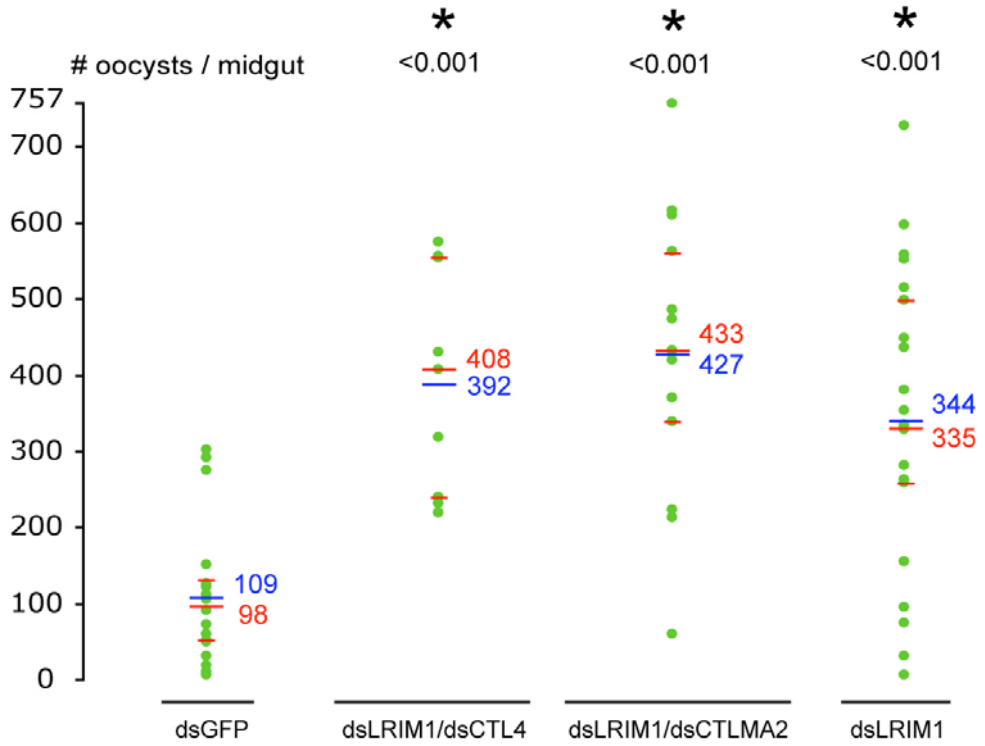


Figure 4.17. Comparison of double to single knockdown experiments of *LRIM1*. Absence of LRIM1 reverts the *CTL4* or *CTLMA2* induced melanising phenotype and produces 4-fold increase in the number of oocysts as compared to the control mosquitoes but not statistically different oocyst numbers from the single *LRIM1* KD mosquitoes. Green dots show fluorescent (living) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisks denote statistically significant difference in pairwise comparisons versus the control mosquitoes (Mann-Whitney test, P-value <0.05. Data have been reprocessed from (Osta et al., 2004))

LRIM1 immunlocalisation in mosquito midguts after Plasmodium infection

To investigate LRIM1 involvement in response to parasite infections in the midgut we performed immunolocalisation experiments 24h after *P. berghei* infections in both S and R mosquitoes (Fig. 4.18). Contrary to the results obtained with immunoblotting experiments (Fig. 4.11), we observed a characteristic and highly reproducible LRIM1 staining in the midgut tissue of the susceptible mosquitoes after parasite infection (Fig 4.18A). LRIM1 is localised in close proximity to the ookinetes and is either surrounding the parasites or is concentrated on the parasite rear end (Fig 4.18A). This localisation was specific to infected mosquitoes and was not observed in the midguts of mosquitoes that have received a non-infectious blood meal (Fig 4.18B). In addition, no signal was detected when the preimmune serum of the antibody was used (data not shown) or when *LRIM1* expression was silenced by dsRNA injection (compare Fig. 4.18C with Fig. 4.18D). Similar localisation was observed in another strain of susceptible mosquitoes (Yaoundé strain, Fig. 4.S8). In

the refractory mosquitoes LRIM1 localisation is also observed near fluorescent ookinetes (Fig. 4.18E) and near newly melanised ookinetes (Fig. 4.18F).

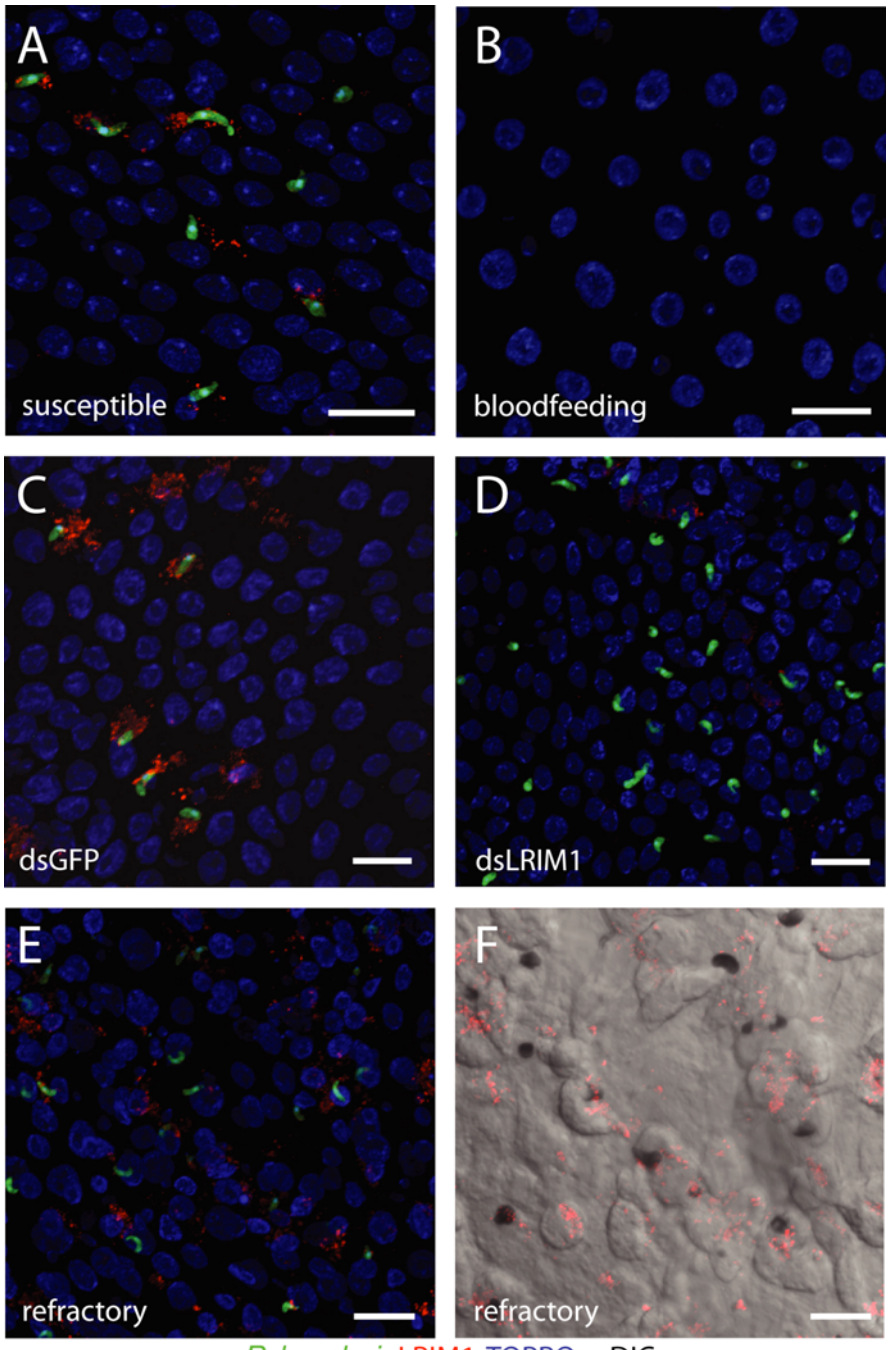


Figure 4.18. LRIM1 localisation in mosquito midguts after *Plasmodium* infection. Confocal stack projections of selected midguts 24hours after infectious (A, C-F) or non-infectious (B) blood meal. LRIM1 is detected in both S (A) and R (E, F) mosquitoes and in dsGFP injected mosquitoes (C) but no LRIM1 is observed in blood fed (B) or dsLRIM1 injected (D) mosquitoes. (scalebar = 20 µm).

We monitored in detail the temporal LRIM1 localisation in mosquito midguts in both S (Fig. 4.19) and R (Fig. 4.20) mosquitoes. In the S mosquitoes, very few parasites are invading the midgut epithelium at 16h p.f. and no LRIM1 localisation is

observed (Fig 4.19A), possibly due to the fact that the mosquito has not yet mounted an immune response against the parasite. LRIM1 is initially detected at 18h p.f., but only near a few ookinetes Its localisation is observed near the majority of the ookinetes only between 20h to 26h p.f. (Fig. 4.19C-F), a period which coincides with the majority of ookinetes invading the midgut epithelium. At 30h post infection, when the majority of the ookinetes have crossed the midgut epithelium and begun transforming into round oocysts, no LRIM1 localisation was detected and this pattern persisted for the remainder of the time periods assessed (Fig. 4.19H,I and 53h p.f. data not shown). Similar observations were made in R mosquitoes (Fig. 4.20). LRIM1 was initially detected at 20h p.i (Fig. 4.20E) and persisted until 30h p.f. (Fig. 4.20G-O). In addition to fluorescent ookinetes, LRIM1 is also found near melanised oocysts (Fig. 4.20G,H, left side). Most probably, those parasites were recently melanised as no LRIM1 staining was detected at later stages of infection (32h p.f.), when all the parasites appeared fully melanised (Fig 4.20Q,R).

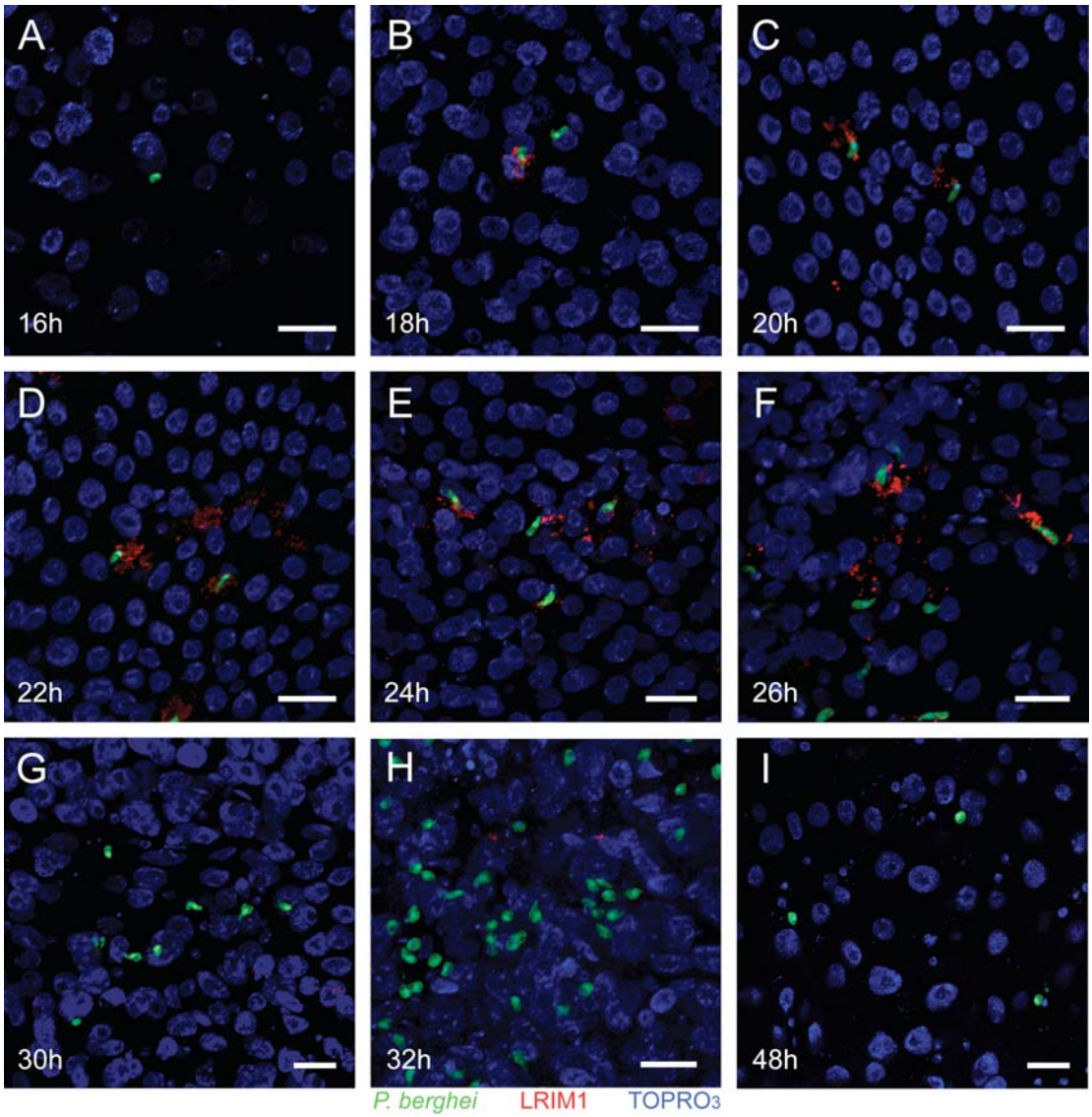


Figure 4.19. LRIM1 localisation in the midguts of S mosquitoes at variable time periods after *P. berghei* infection. All pictures are confocal stack projections of selected midguts at various time periods after infection (scalebar = 20 μm).

The small difference in the temporal localisation of LRIM1 in the S and the R strains can be attributed to differences in the intensity of infection. Indeed, additional experiments showed that the appearance and disappearance of LRIM1 may vary slightly (18h-20h and 30-32h respectively, data not shown). Nevertheless, these results indicate that LRIM1 staining in the midgut upon *P. berghei* infections is specific to the ookinete stage of the parasite and ranges from the start of the invasion until the time that the majority of the ookinetes have invaded the mosquito midgut. These ookinetes are either transforming into immotile oocysts in the S or are melanised in the space between the basal lamina and the basal labyrinth in the R mosquito strain.

LRIM1 localisation in the midgut epithelium was analysed in more detail. To distinguish between the midgut lumen (apical side), which is in contact with the infectious blood meal and the basal side which is bathed in hemolymph, a phalloidin-conjugated fluorophore was used. Phalloidin attaches to the actin molecules of muscle cells lining the basal side of the midgut epithelium. Again, LRIM1 was observed to surround the fluorescent parasite (Fig. 4.21A). Examination of individual planes of the confocal stack revealed that LRIM1 was proximal to the parasite in the basal side and became more distal to the parasite during the transition to the lumen side (Fig. 4.21C,E,G). Fig. 4.21H shows a characteristic cell expulsion towards the lumen side of the epithelium, indicating the induction of apoptosis after parasite invasion; interestingly, LRIM1 is mostly confined to this cell. It is however, unclear, if LRIM1 localisation precedes or follows apoptosis. Immunohostochemical experiments performed in the absence of detergent, which presumably allow the antibodies to access the cell interior, suggested that LRIM1 is intracellular, as no LRIM1 staining was detected (Fig. 4.S9). Whether LRIM1 is expressed by midgut cells upon midgut invasion or is secreted by hemocyte cells, endocytosed by cells and recruited in the invaded epithelium remains to be elucidated. Anothet possible explanation is that in the absense of detergent, LRIM1 protein is involved in protein-protein interactions, yielding the peptide inaccessible to the LRIM1 antibodies.

A careful analysis of LRIM1 localisation reveals that there are areas where no fluorescent (living) parasites are detected (e.g. Fig 4.18C, Fig. 4.20D, Fig. 4.21). Possibly, this LRIM1 localisation is due to the existence of parasites that have previously been killed leading to absense of fluorescence. A similar phenomenon has been observed for another protein, TEP1 (Blandin et al., 2004). Future immulolocalisation experiments with parasite specific membrane antbodies, e.g. antibodies against the membrane protein P28, will show whether LRIM1 localisation is due to the presence of killed parasites.

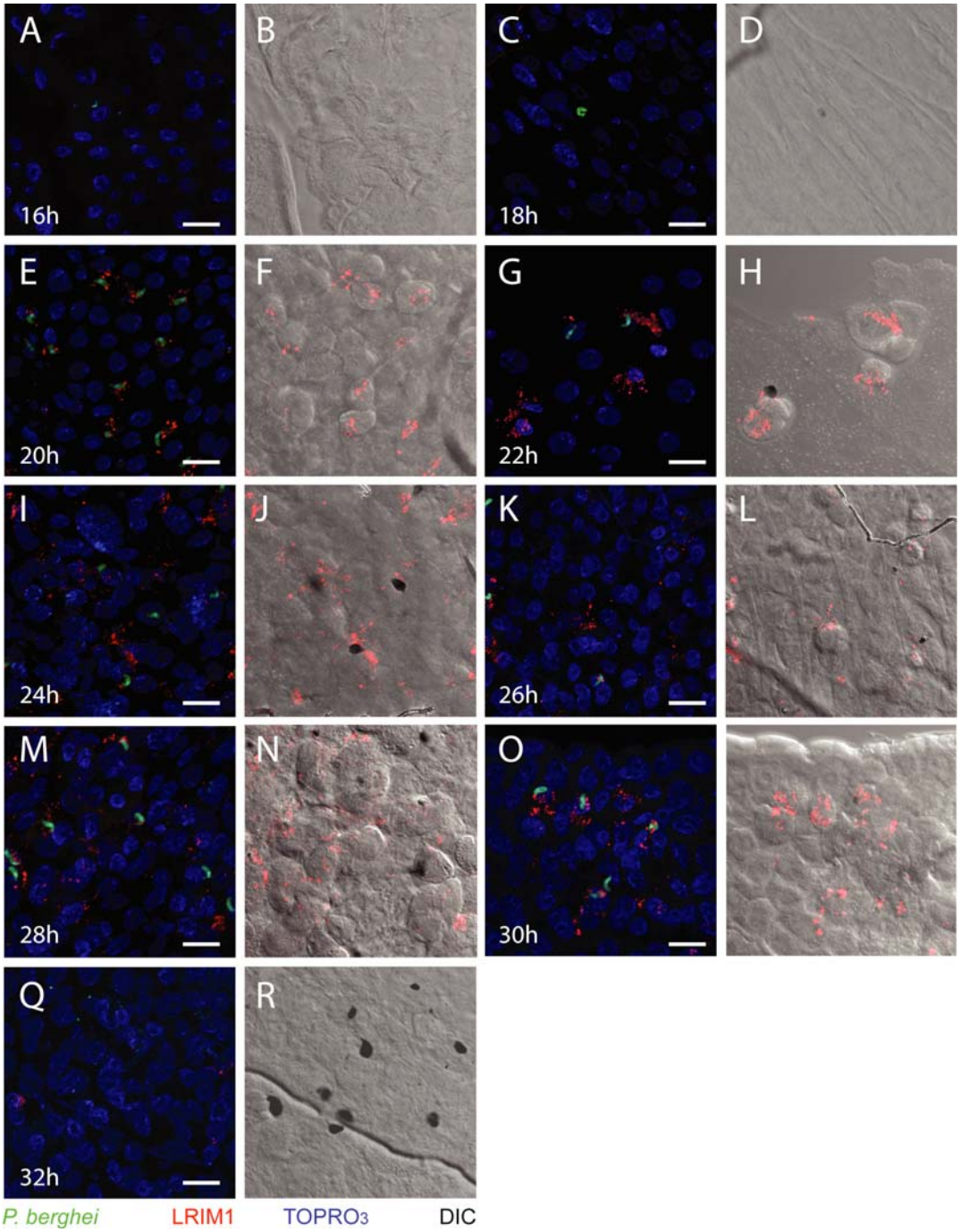


Figure 4.20. LRIM1 localisation in R mosquitoes at various time periods after *P. berghei* infection. All pictures are confocal stack projections of in mosquito midguts (scalebar = 20 μm).

In the R mosquitoes, LRIM1 was located predominantly in cells expulsing from the midgut epithelium due to parasite invasion (Fig.4.22A). Individual planes of the confocal stack detected LRIM1 near the melanised parasite in the basal side of the epithelium (Fig 4.22B) and its localisation was progressively becoming distal during the transition from the basal to the lumen side of the midgut (Fig 4.22C-E). Our results indicate that LRIM1 follows the parasite movement in the interior of the midgut cell compartment, starting from parasite entry in the lumen side until its exit

to the extracellular space at the basal side. However, in the absence of real-time observations of LRIM1 localisation in proximity to the parasite, such hypothesis remains elusive.

In conclusion, we observed a specific and reproducible localisation of LRIM1 at the vicinity of both fluorescent and newly melanised parasites and we showed that LRIM1 localisation is specific to the ookinete stage. Importantly, although co-localisation of LRIM1 with the parasite is not exact, LRIM1 is definitely induced by the parasite, suggesting that the LRIM1 function in parasite killing and melanisation is mediated indirectly, for example by interaction of the protein with other interacting partners.

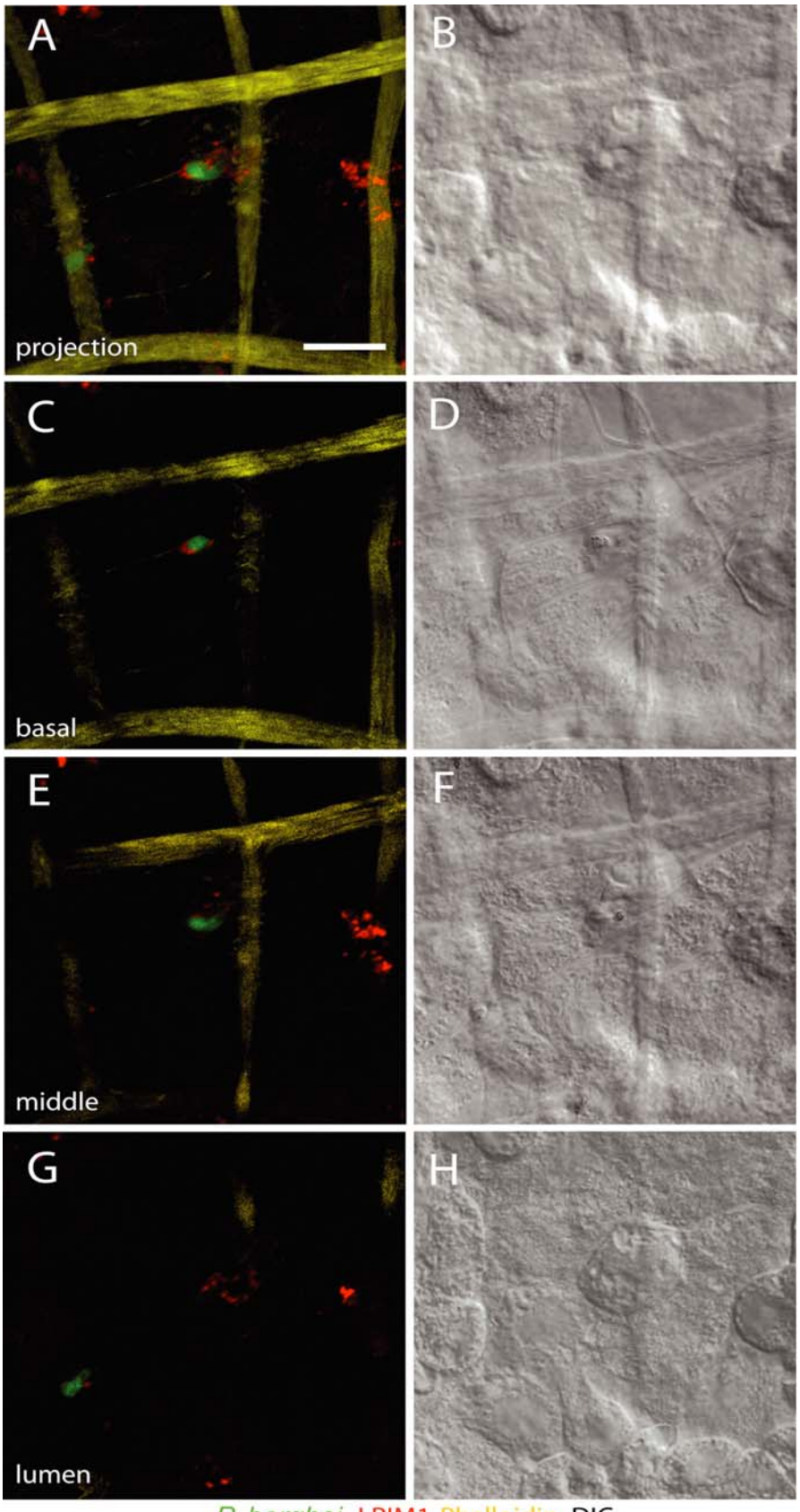


Fig. 4.21. LRIM1 localisation in the S strain of *A. gambiae*. Pictures represent confocal stack projections (A, B) and individual sections (C-H) of mosquito midguts. Notice that LRIM1 staining is apparent in all sections of the midgut. (scalebar = 20 μm).

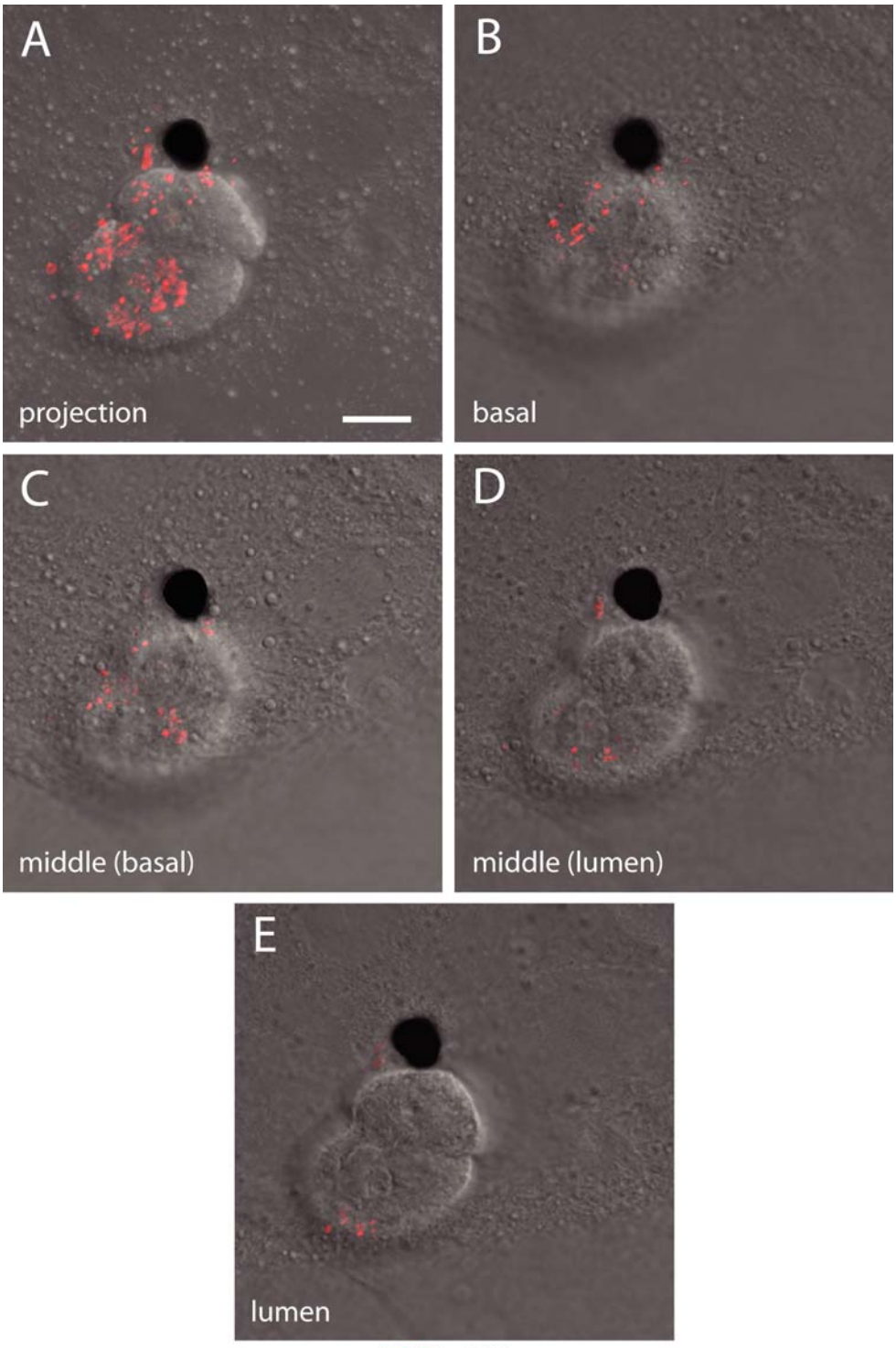


Figure 4.22. LRIM1 expression in the R strain of *A.gambaie*. Pictures represent confocal stack projections (A) and individual sections (B-E) of DIC and fluorescent pictures. LRIM1 colocalises in the invaded epithelial cells and in the vicinity of the melanised parasite. In addition, a proximal to distal staining of LRIM1 relative to the parasite is observed from the basal to the lumen side. (scalebar = 10 μM).

Discussion

LRR proteins and their involvement in innate immunity

During the journey of the malaria parasite through the mosquito many parasite losses occur. Undoubtedly, the crossing of the midgut epithelium is one of the most critical bottlenecks. Studies have shown that this bottleneck can reduce the number of parasites to single digit numbers (Sinden, 1999). However, even if one parasite survives and transforms into an oocyst, it will multiply to produce thousands of sporozoites capable to infect another vertebrate host. A number of studies have implicated mosquito innate immune responses for, at least part, of these parasite losses. To this end, our laboratory has initiated a comprehensive study of innate immunity in the mosquito. Current research aims to characterise parasite responses for known innate immune molecules and to identify new molecules involved in such responses.

The leucine rich repeat family contains proteins that are involved in a variety of different functions. Toll in *Drosophila* and Toll-like receptors are examples of LRR proteins with central roles in pathways for the elimination of pathogens. Until recently, no LRR-containing protein was implicated in *Plasmodium* parasite development. Our study describes such a protein, termed *leucine rich repeat immune gene, LRIM1*, which was discovered in a large scale transcriptomic analysis.

LRR domain characterisation and expression

LRIM1 domain analysis indicated the presence of two generic protein-protein interaction domains: an LRR domain at the 5' end and a coiled-coil domain at the 3' end. This generic domain architecture suggests interaction of LRIM1 with other molecules and may also explain the inability to detect orthologues of this protein to other organisms; only weak similarity to other LRR domains was detected. Probably, small variations in the peptide sequences in these protein-interaction domains are contributing to the three-dimensional structure and confer specificity to the target protein. Consequently, the remainder of the protein sequence can be variable, thus displaying no detectable similarity to any known protein. Interestingly, molecular modelling of the protein indicated that the 'oddie' peptide, in the middle of the LRR sequence, might contribute a smaller, albeit proper, LRR repeat or separate the LRR

domain into two parts. Perhaps a crystallographic study of the LRR domain might shed more light about the structure of the repeats and the 'oddie' peptide, as well as provide hints about possible interaction partners.

Several lines of evidence suggest that LRIM1 is primarily of hemocyte origin. LRIM1 includes a transmembrane or signal peptide domain, indicating it is either membrane-bound or secreted. A comprehensive study of RNA abundance in mosquito tissues also noted overexpression of LRIM1 in circulating hemocytes compared to head and carcass tissues. Finally, protein blots assays clearly detected LRIM1 in the mosquito hemolyph. LRIM1 expression in other tissues (mainly midgut epithelial cells and carcasses) remains an open question. RNA abundance studies indicated presence of LRIM1 transcripts in midguts and carcasses, especially after *Plasmodium* infection. The presence of LRIM1 RNA is possibly due to circulating hemocytes which are co-isolated with those tissues. However, LRIM1 protein is detected in midgut tissues almost immediately after invasion of parasites and persists for the entire time period of ookinete invasion. In light of this protein localisation, *LRIM1* upregulation in midguts and carcasses at 24h post infection is possibly due to a feedback mechanism, aiming to replenish LRIM1 protein molecules that are rapidly recruited to the mosquito midguts. The absence of hemolymp constitutive protein marker prevents precise quantification and comparison of LRIM1 protein levels at variable time periods after parasite infection. Recently, however, a serine protease inhibitor, SRPN2, was shown to be constitutively expressed upon parasite infection (Michel, K., personal communication) and further experiments will facilitate precise LRIM1 protein quantification in the near future.

LRIM1 localisation in the midgut cells show a characteristic granular appearance and suggest that this specific localisation may be the result of interaction with other protein factors. A similar localisation has been described for the P28 membrane protein of *P. berghei*, which sheds P28 protein molecules in ookinete cultures (Sinden et al., 1987). It would be interesting to compare the pattern of LRIM1 localisation with P28 and possibly other parasite membrane proteins, to investigate possible interaction of LRIM1 with the parasite. In addition, P28 co-localisation experiments will help determined the proportion amount of parasites (living or killed) that contain LRIM1 localisation.

LRIM1 in P. berghei immunity

The development of the RNAi technique in adult mosquitoes opened unprecedented opportunities to conduct gene functional studies. Using this technique we assessed *LRIM1* function during *P. berghei* parasite infection. Initial results indicated a large increase in the number of developing oocysts in both the S and R strains of mosquitoes. This increase, however, varied in individual experiments, indicating that additional factors may influence parasite numbers.

A factor that has not been addressed is the intensity of parasite infection. Results from multi-factorial analyses (two-way ANOVA in the pairwise comparisons and three-way ANOVA in the comparative experiments, data not shown) indicated that there is significant variability between the individual experiments. We propose that low infections may not be sufficient to demonstrate a significant difference between control and *dsLRIM1* mosquitoes. Monitoring the number of ingested gametocytes in future infection experiments will better define the relationship of parasite concentration and *LRIM1*-mediated immune responses resulting in parasite killing.

Interestingly, absence of *LRIM1* from the R mosquitoes abolishes the melanisation phenotype. Melanisation is a general mechanism for the enclosure of pathogenic organisms in a polymer composed of melanin. In *A. gambiae*, the refractory strains that melanises the *P. berghei* parasites has been established by genetic selection for refractoriness from the susceptible strain of mosquitoes (Collins et al., 1986). Our intra-strain comparison showed that in LRIM1 KD results in equal fluorescent (living) parasite numbers in the S and R strains. This result suggests that the complex trait of refractoriness could, at least partly, be explained by *LRIM1*. In addition, we observed an effect of *LRIM1* epistatic to the action of *CTL4* and *CTLMA2*, which appear to protect parasites against melanisation in the R mosquitoes. The double knockdown experiment of LRIM1 and the C-type lectins is the first example of genetic epistasis that highlighted the presence of agonists (like C-type lectins) and antagonists (like *LRIM1*) in parasite development (Osta et al., 2004).

Indication of interaction with other immune related genes

In an effort to define LRIM1 function in relation to parasite development, genetic and/or physical interaction with other proteins is needed. Recently, another LRR protein showed a similar phenotype to *LRIM1* in parasite killing in the S mosquito

strain (Riehle et al., 2006) and experiments will determine if those proteins share functional similarity. Interstingly, a similar phenotype to LRIM1 parasite killing and melanisation has also been reported for a thioester containing protein, TEP1 (Blandin et al., 2004; Levashina et al., 2001). This strikingly similarity suggested that LRIM1 and TEP1 might be implicated in the same functional pathway. Initial double KD experiments between TEP1 and LRIM1 (data not shown) did not show significant difference when compared to the single KD experiments. In the absense of LRIM1, TEP1 abolishes its binding to *P. berghei* parasites, suggesting a possible downstream role of TEP1 to the function of LRIM1. Whether absence of TEP1 affects LRIM1 localisation is currently being investigated. In addition, the striking developmental expression similarity between LRIM1 and several serine proteases suggests possible functional relevance. Future co-localisation experiments of *LRIM1* with *TEP1* and serine proteases will therefore help to establish links with other known immune related molecules and may provide further insight to the mechanism of *LRIM1* function.

LRIM1 in P. falciparum immunity

The experiments reported in this thesis were conducted using the laboratory model parasite, *P. berghei*, which, however, is not responsible for human malaria. It is interesting to assess whether LRIM1 is able to mount an immune response against human malaria parasites and, in particular, in natural parasite populations in the sub-Saharan Africa. Experiments with natural populations of *P. falciparum* from Cameroon indicated no difference in the oocyst numbers between control and *dsLRIM1* mosquitoes. While the possibility that *P. falciparum* is recognised by a different mechanism than *P. berghe* can not be excluded*i*, which is independent of *LRIM1*-mediated immune responses, in light of the intra-strain comparison results, an alternative explanation may also be possible: *P. falciparum* infections are significantly lower than *P. berghei* and thus, low infectivity, might not allow for a measurable difference in oocyst numbers. Experiments in laboratory infections of *P. falciparum* are in progress, in order to establish if *P. falciparum* infections are characterised by similar LRIM1 protein localisation upon parasite invasion.

LRIM1 in bacterial immunity

A growing number of evidence also suggest *LRIM1* involvement in bacterial immune responses. *LRIM1* is involved in phagocytosis against *E. coli*, but not *S. aureus* (Moita et al., 2005). Our data demonstrated a concentration dependent effect of LRIM1 in *E. coli* infections and a general effect in *S. aureus* infections irrespective of concentration. However, in the latter bacterium a concentration-dependent effect might be observed using a significantly smaller bacterial concentration that in this study. Perhaps, LRIM1 is also implicated in bacterial immune responses other than phagocytosis, as LRIM1 is downregulated in the absence of REL2 (Meister et al., 2005). Future experiments with different bacteria species aims to better define the role of LRIM1 in bacterial innate immunity.

Conclusion

In conclusion, *LRIM1* has an important role in killing and melanisation of the malaria parasites, as well as innate immune responses against bacteria. Future studies will focus on the identification of putative interactive partners. Those studies will shed light to LRIM1 mechanism of function and may contribute towards a greater understanding of innate immune responses, which can potentially be used for future malaria interventions in sub-Saharan Africa.

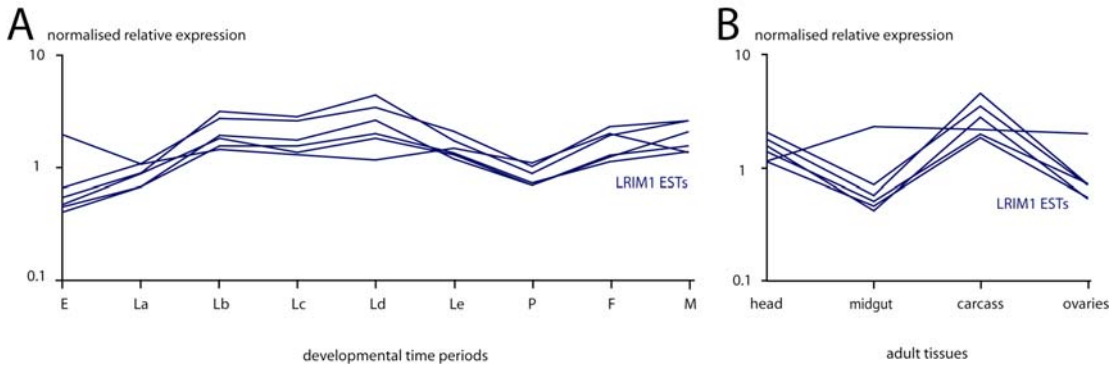


Figure 4.S1. Expression of the ESTs of LRIM1 in the A) developmental life cycle and the B) adult tissues. Notice that one of the ESTs shows differential expression in embryos and another one in the adult tissues experiment

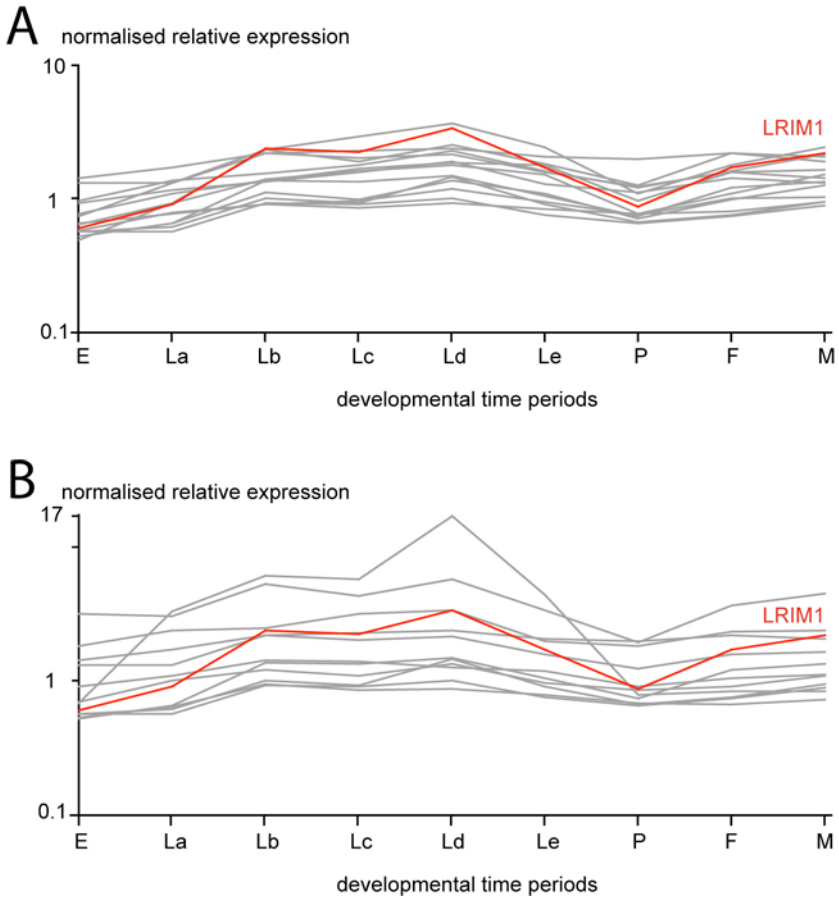


Figure 4.S2. Expression profiles of the TCLAG contigs displaying expression similar to LRIM1. TCLAG contigs showing above 0.9 similarity in A) Pearson and B) Spearman correlation coefficients have been plotted. Red lines show LRIM1 expression and relative expression is in log scale.

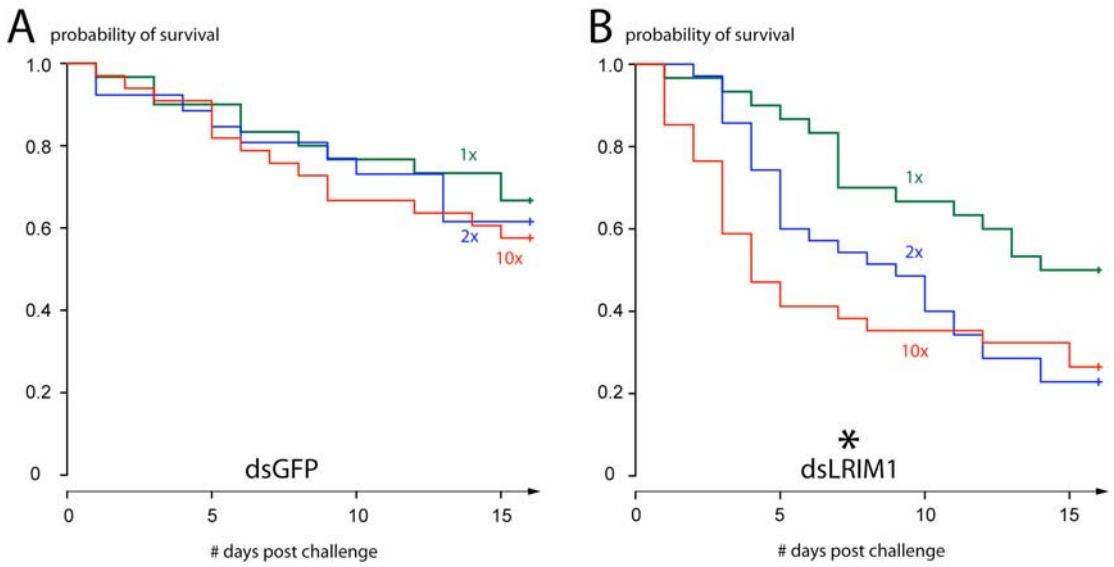


Figure 4.S3. Survival of control and *dsLRIM1* knockdown mosquitoes after variable concentrations of *E. coli*. Kaplan- Meier curves of the A) *dsGFP* and the B) *dsLRIM1* mosquitoes were drawn together and asterisks denote statistical significance. Notice that while increasing concentrations of bacteria did not have an effect on dsGFP mosquitoes (P-value = 0.737) they showed significant and concentration dependent decrease in the survival probability in the *dsLRIM1* injected mosquitoes (P-value = 0.0184)

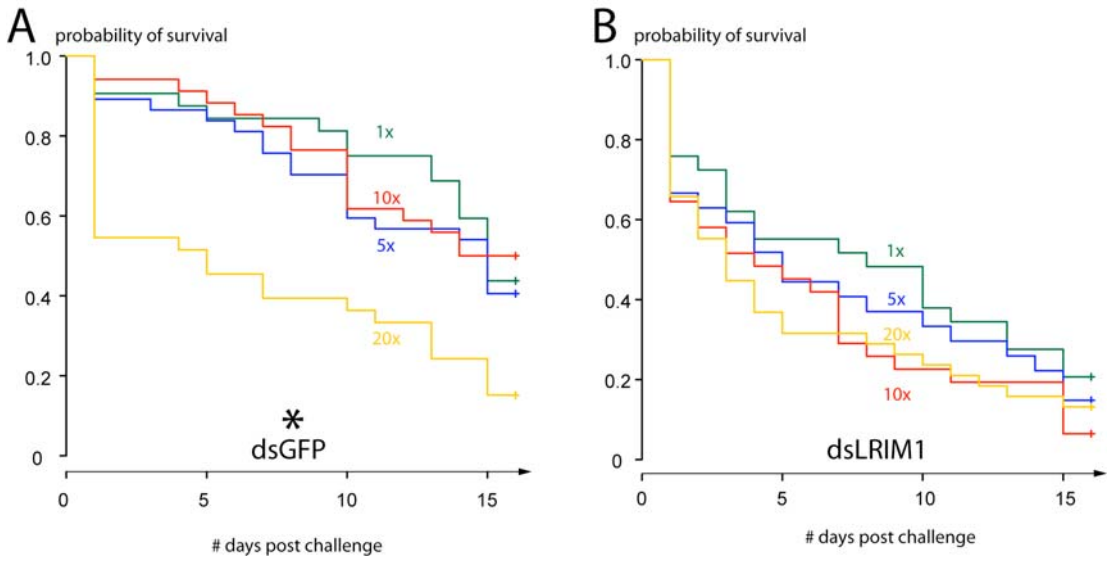


Figure 4.S4. Survival of control and *dsLRIM1* knockdown mosquitoes after variable concentrations of *S. aureus*. Kaplan- Meier curves of the A) *dsGFP* and the B) *dsLRIM1* mosquitoes have been drawn together and asterisks denote statistical significance. Notice that survival the *dsGFP* injected is similar in the 1×, 2×, 10× concentrations (P-value = 0.786) but differs in 20× concentration (P-value < 0.001). Conversely, the *dsLRIM1* mosquitoes show similar survival curves (P-value = 0.374) in all the concentrations tested.

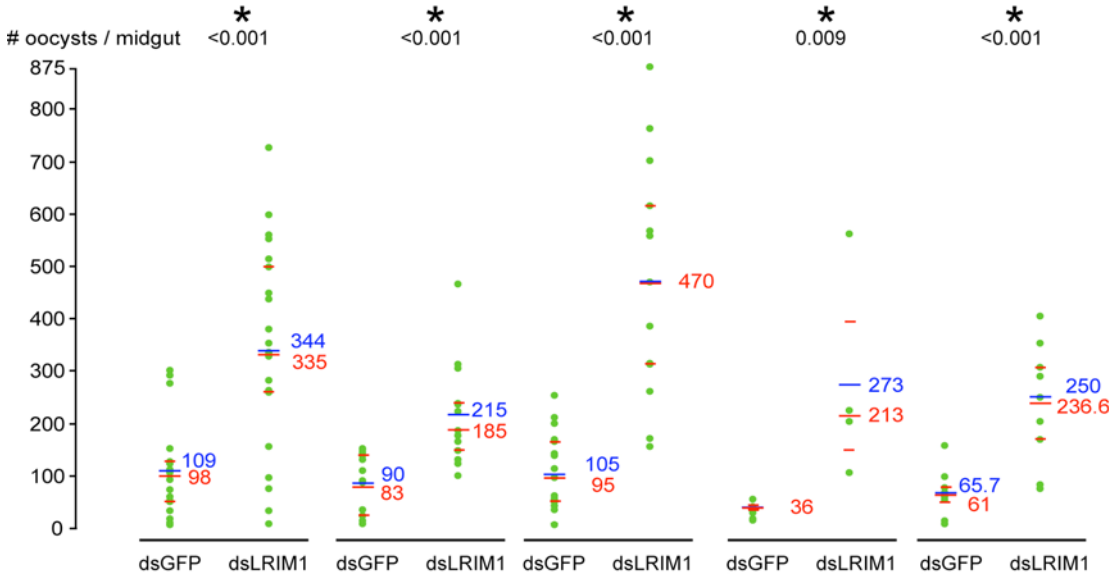


Figure 4.S5. Effects of *LRIM1* KD in oocysts numbers in the individual experiments in the susceptible mosquitoes. Variable increase in the oocyst numbers was observed in the individual 5 experiments. Green dots show fluorescent (living) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisks denote statistically significant difference between control and *dsLRIM1* mosquitoes (Mann-Whitney test, P-value <0.05) and P-values are indicated.

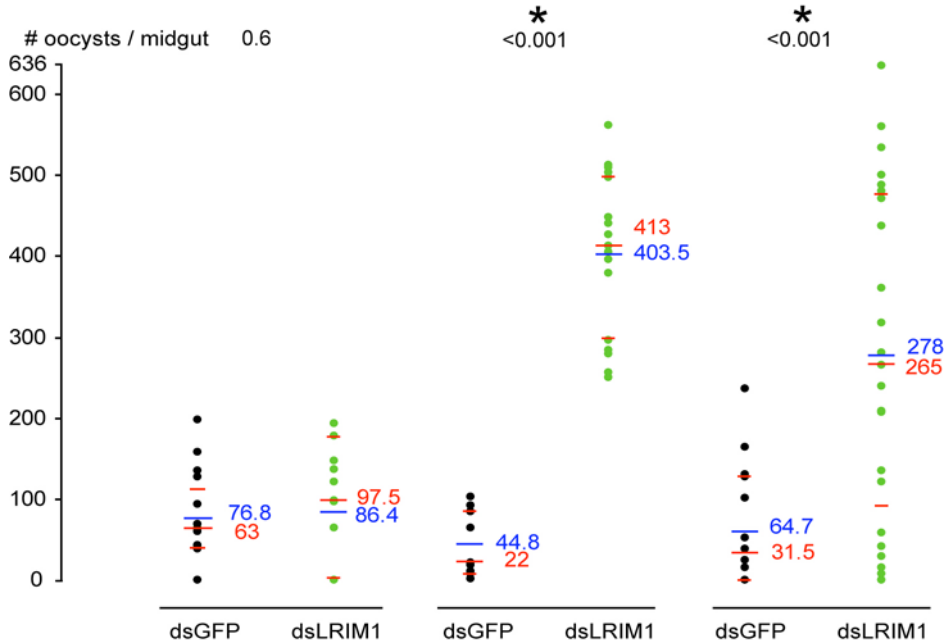


Figure 4.S6. Effects of *LRIM1* KD in oocysts numbers in the individual experiments in the refractory mosquitoes. No increase in parasite numbers was detected on the first experiment whereas significant increase was detected in the remaining two experiments. Green dots show fluorescent (living) parasites, black dots melanised (killed) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisks denote statistically significant difference between control and *dsLRIM1* mosquitoes (Mann-Whitney test, P-value <0.05) and P-values are indicated.

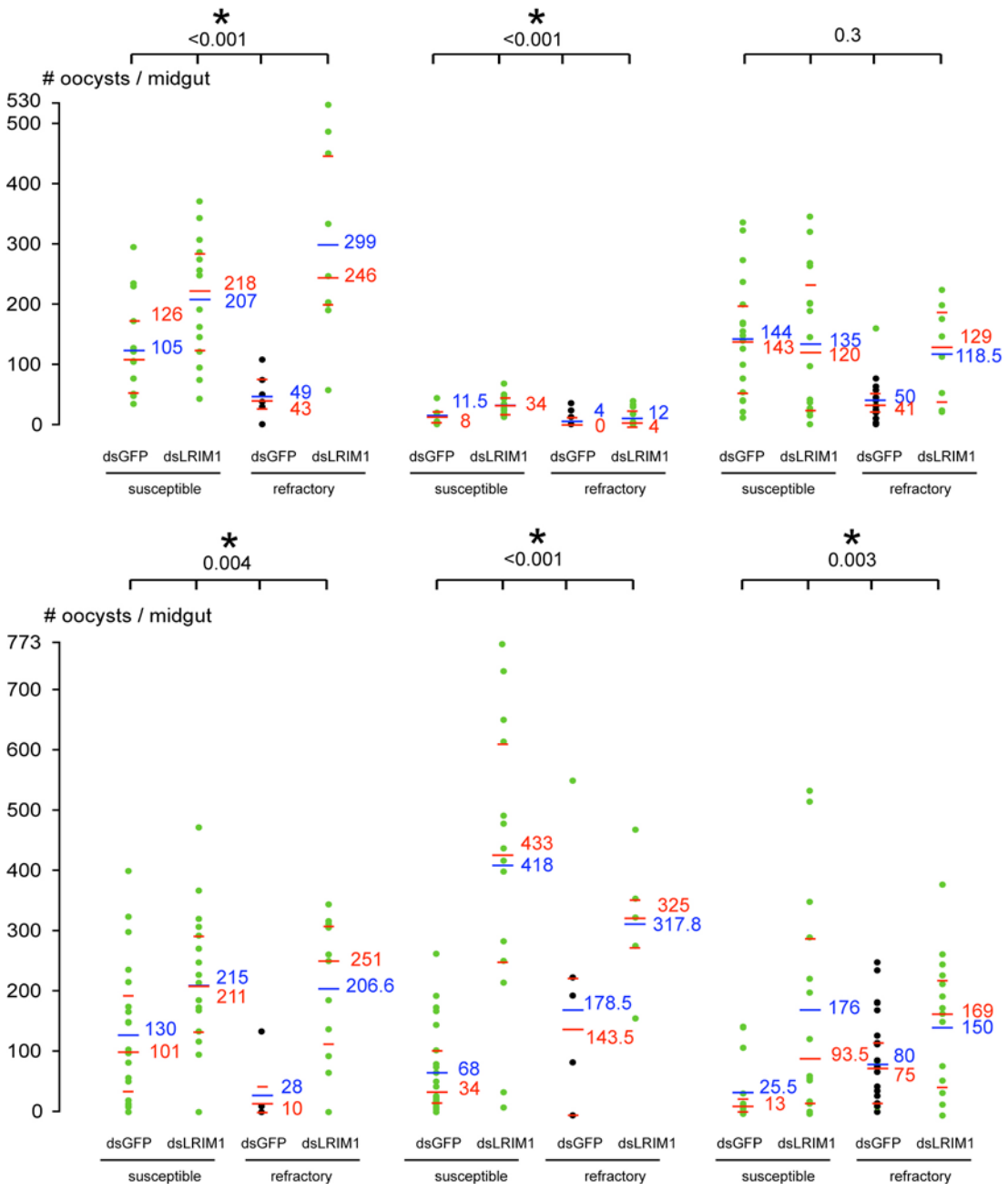


Figure 4.S7. Comparison of oocyst numbers in susceptible and refractory mosquitoes in individual experiments infected with equal parasite numbers. Green dots show fluorescent (living) parasites, black dots melanised (killed) parasites, blue lines show means, long red lines show medians and short red lines show the 25% and 75% quartiles. Asterisk denotes statistically significant difference between strains and control and dsLRIM1 mosquitoes (Kruskal-Wallis test, P-value < 0.001). and P-valies are indicated.

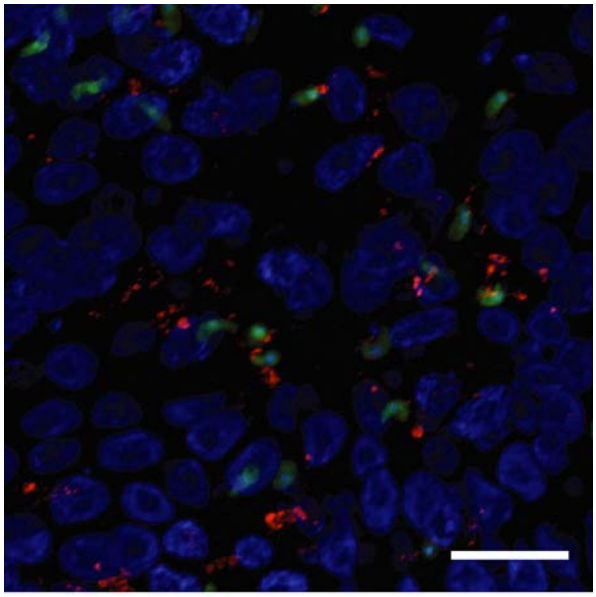


Figure 4.S8. LRIM1 localisation in the Yaoundé strain, an additional susceptible mosquito. Selected area of a mosquito midgut confocal stack projection 24h after infection with *P. berghei* parasites. A similar pattern of LRIM1 staining to the susceptible G3 strain has been observed. (scalebar = 20 μm)

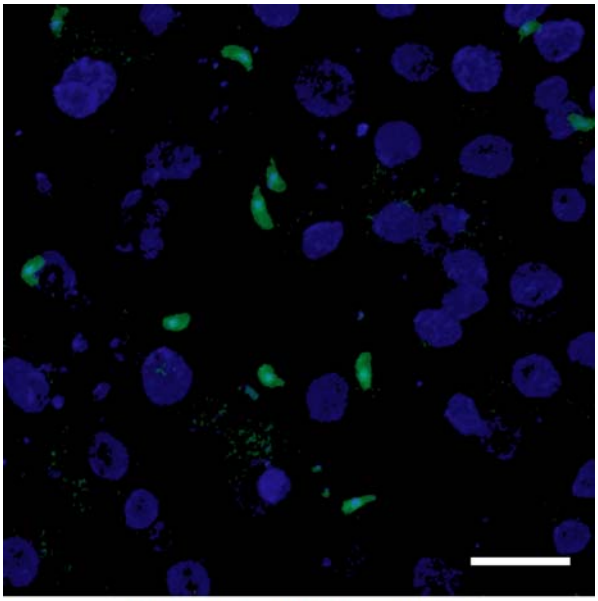


Figure 4.S9. Evidence for intracellular localisation of LRIM1. Confocal stack projection of a selected area of midgut of a susceptible mosquito at 24h after parasite infection. In the absence of detergent, the LRIM1 antibodies are not able to access the intracellular space and thus, no LRIM1 localisation is detected. (scalebar = 20 μm).

Bibliography

Abraham, E. G., Pinto, S. B., Ghosh, A., Vanlandingham, D. L., Budd, A., Higgs, S., Kafatos, F. C., Jacobs-Lorena, M., and Michel, K. (2005). An immune-responsive serpin, SRPN6, mediates mosquito defense against malaria parasites. Proc Natl Acad Sci U S A *102*, 16327-16332.

Abramoff, M. D., Magelhaes, P. J., and Ram, S. J. (2004). Image processing with ImageJ. Biophotonics International *11*, 36-42.

Aderem, A., and Underhill, D. M. (1999). Mechanisms of phagocytosis in macrophages. Annu Rev Immunol *17*, 593-623.

Amino, R., Menard, R., and Frischknecht, F. (2005). In vivo imaging of malaria parasites--recent advances and future directions. Curr Opin Microbiol *8*, 407-414.

Amino, R., Thiberge, S., Martin, B., Celli, S., Shorte, S., Frischknecht, F., and Menard, R. (2006). Quantitative imaging of Plasmodium transmission from mosquito to mammal. Nat Med *12*, 220-224.

Anderson, K. V., Bokla, L., and Nusslein-Volhard, C. (1985). Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product. Cell *42*, 791-798.

Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. J Struct Biol *134*, 117-131.

Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol *298*, 521-537.

Angel, J. L. (1966). Porotic hyperostosis, anemias, malarias, and marshes in the prehistoric Eastern Mediterranean. Science *153*, 760-763.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res *32*, D115-119.

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of Drosophila melanogaster. Science *297*, 2270-2275.
Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Asling, B., Dushay, M. S., and Hultmark, D. (1995). Identification of early genes in the Drosophila immune response by PCR-based differential display: the Attacin A gene and the evolution of attacin-like proteins. Insect Biochem Mol Biol *25*, 511-518.

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., Bumgarner, R. E., Bushel, P. R., Chaturvedi, K., Choi, D., *et al.* (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods *2*, 351-356.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al.* (2004). The Pfam protein families database. Nucleic Acids Res *32*, D138-141.

Bates, M. (1949). The natural history of mosquitoes (New York,: Macmillan).
Baton, L. A., and Ranford-Cartwright, L. C. (2005). How do malaria ookinetes cross the mosquito midgut wall? Trends Parasitol *21*, 22-28.

Beier, J. C. (1998). Malaria parasite development in mosquitoes. Annu Rev Entomol *43*, 519-543.

Bell, J. K., Mullen, G. E., Leifer, C. A., Mazzoni, A., Davies, D. R., and Segal, D. M. (2003). Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol *24*, 528-533.

Bellotto, M., Bopp, D., Senti, K. A., Burke, R., Deak, P., Maroy, P., Dickson, B., Basler, K., and Hafen, E. (2002). Maternal-effect loci involved in Drosophila oogenesis and embryogenesis: P element-induced mutations on the third chromosome. Int J Dev Biol *46*, 149-157.

Belyakin, S. N., Christophides, G. K., Alekseyenko, A. A., Kriventseva, E. V., Belyaeva, E. S., Nanayev, R. A., Makunin, I. V., Kafatos, F. C., and Zhimulev, I. F. (2005). Genomic analysis of Drosophila chromosome underreplication reveals a link between replication control and transcriptional territories. Proc Natl Acad Sci U S A *102*, 8269-8274.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res *32*, D23-26.

Bergelson, J., Kreitman, M., Stahl, E. A., and Tian, D. (2001). Evolutionary dynamics of plant R-genes. Science *292*, 2281-2285.

Biessmann, H., Walter, M. F., Dimitratos, S., and Woods, D. (2002). Isolation of cDNA clones encoding putative odourant binding proteins from the antennae of the malaria-transmitting mosquito, Anopheles gambiae. Insect Mol Biol *11*, 123-132.

Billker, O., Lindo, V., Panico, M., Etienne, A. E., Paxton, T., Dell, A., Rogers, M., Sinden, R. E., and Morris, H. R. (1998). Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. Nature *392*, 289-292.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., *et al.* (2004). An overview of Ensembl. Genome Res *14*, 925-928.

Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res *12*, 739-748.

Blandin, S., Moita, L. F., Kocher, T., Wilm, M., Kafatos, F. C., and Levashina, E. A. (2002). Reverse genetics in the mosquito Anopheles gambiae: targeted disruption of the Defensin gene. EMBO Rep *3*, 852-856.

Blandin, S., Shiao, S. H., Moita, L. F., Janse, C. J., Waters, A. P., Kafatos, F. C., and Levashina, E. A. (2004). Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. Cell *116*, 661-670.

Bonaldo, M. F., Lennon, G., and Soares, M. B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res *6*, 791-806.

Breman, J. G. (2001). The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. Am J Trop Med Hyg *64*, 1-11.

Bulet, P., Dimarcq, J. L., Hetru, C., Lagueux, M., Charlet, M., Hegy, G., Van Dorsselaer, A., and Hoffmann, J. A. (1993). A novel inducible antibacterial peptide of Drosophila carries an O-glycosylated substitution. J Biol Chem *268*, 14893-14897.

Bulet, P., Hetru, C., Dimarcq, J. L., and Hoffmann, D. (1999). Antimicrobial peptides in insects; structure and function. Dev Comp Immunol *23*, 329-344.

Cabrera, C. V., Alonso, M. C., and Huikeshoven, H. (1994). Regulation of scute function by extramacrochaete in vitro and in vivo. Development *120*, 3595-3603.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res *32*, D262-266.

Catteruccia, F., Nolan, T., Loukeris, T. G., Blass, C., Savakis, C., Kafatos, F. C., and Crisanti, A. (2000). Stable germline transformation of the malaria mosquito Anopheles stephensi. Nature *405*, 959-962.

Cha, B. J., Koppetsch, B. S., and Theurkauf, W. E. (2001). In vivo analysis of Drosophila bicoid mRNA localization reveals a novel microtubule-dependent axis specification pathway. Cell *106*, 35-46.

Chamaillard, M., Girardin, S. E., Viala, J., and Philpott, D. J. (2003). Nods, Nalps and Naip: intracellular regulators of bacterial-induced inflammation. Cell Microbiol *5*, 581-592.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell *2*, 65-73.

Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. Nat Genet *27*, 48-54.

Choe, K. M., Werner, T., Stoven, S., Hultmark, D., and Anderson, K. V. (2002). Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in Drosophila. Science *296*, 359-362.

Christophers, S. R. (1960). Aëdes aegypti (L.), the yellow fever mosquito; its life history, bionomics, and structure (Cambridge [Eng.]: University Press).

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P. T., Collins, F. H., Danielli, A., Dimopoulos, G., *et al.* (2002). Immunity-related genes and gene families in Anopheles gambiae. Science *298*, 159-165.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Clements, A. N., and Clements, A. N. (1992). The biology of mosquitoes, 1st edn (London ; New York: Chapman & Hall).

Collins, F. H., and Paskewitz, S. M. (1995). Malaria: current and future prospects for control. Annu Rev Entomol *40*, 195-219.

Collins, F. H., Sakai, R. K., Vernick, K. D., Paskewitz, S., Seeley, D. C., Miller, L. H., Collins, W. E., Campbell, C. C., and Gwadz, R. W. (1986). Genetic selection of a Plasmodium-refractory strain of the malaria vector Anopheles gambiae. Science *234*, 607-610.

Dana, A. N., Hong, Y. S., Kern, M. K., Hillenmeyer, M. E., Harker, B. W., Lobo, N. F., Hogan, J. R., Romans, P., and Collins, F. H. (2005). Gene expression patterns associated with blood-feeding in the malaria mosquito Anopheles gambiae. BMC Genomics *6*, 5.

Daniel, J. A., Torok, M. S., Sun, Z. W., Schieltz, D., Allis, C. D., Yates, J. R., 3rd, and Grant, P. A. (2004). Deubiquitination of histone H2B by a yeast acetyltransferase complex regulates transcription. J Biol Chem *279*, 1867-1871.

David, J. P., Strode, C., Vontas, J., Nikou, D., Vaughan, A., Pignatelli, P. M., Louis, C., Hemingway, J., and Ranson, H. (2005). The Anopheles gambiae detoxification chip: a highly specific microarray to study metabolic-based insecticide resistance in malaria vectors. Proc Natl Acad Sci U S A *102*, 4080-4084.

De Gregorio, E., Han, S. J., Lee, W. J., Baek, M. J., Osaki, T., Kawabata, S., Lee, B. L., Iwanaga, S., Lemaitre, B., and Brey, P. T. (2002). An immune-responsive Serpin regulates the melanization cascade in Drosophila. Dev Cell *3*, 581-592.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet *14*, 457-460.

Dimarcq, J. L., Hoffmann, D., Meister, M., Bulet, P., Lanot, R., Reichhart, J. M., and Hoffmann, J. A. (1994). Characterization and transcriptional profiles of a Drosophila gene encoding an insect defensin. A study in insect immunity. Eur J Biochem *221*, 201-209.

Dimopoulos, G. (2003). Insect immunity and its implication in mosquito-malaria interactions. Cell Microbiol *5*, 3-14.

Dimopoulos, G., Casavant, T. L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., Schultz, J., Benes, V., Bork, P., Ansorge, W., *et al.* (2000). Anopheles gambiae pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. Proc Natl Acad Sci U S A *97*, 6619-6624.

Dimopoulos, G., Christophides, G. K., Meister, S., Schultz, J., White, K. P., Barillas-Mury, C., and Kafatos, F. C. (2002). Genome expression analysis of Anopheles gambiae: responses to injury, bacterial challenge, and malaria infection. Proc Natl Acad Sci U S A *99*, 8814-8819.

Dimopoulos, G., Richman, A., Muller, H. M., and Kafatos, F. C. (1997). Molecular immune responses of the mosquito Anopheles gambiae to bacteria and malaria parasites. Proc Natl Acad Sci U S A *94*, 11508-11513.

Dimopoulos, G., Seeley, D., Wolf, A., and Kafatos, F. C. (1998). Malaria infection of the mosquito Anopheles gambiae activates immune-responsive genes during critical transition stages of the parasite life cycle. Embo J *17*, 6115-6123.

Drysdale, R. A., and Crosby, M. A. (2005). FlyBase: genes and gene models. Nucleic Acids Res *33*, D390-395.

Durbin, R. (1998). Biological sequence analysis : probabalistic models of proteins and nucleic acids (Cambridge, UK New York: Cambridge University Press).

Dushay, M. S., Asling, B., and Hultmark, D. (1996). Origins of immunity: Relish, a compound Rel-like gene in the antibacterial defense of Drosophila. Proc Natl Acad Sci U S A *93*, 10343-10347.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Fearon, D. T., and Locksley, R. M. (1996). The instructive role of innate immunity in the acquired immune response. Science *272*, 50-53.

Fehlbaum, P., Bulet, P., Michaut, L., Lagueux, M., Broekaert, W. F., Hetru, C., and Hoffmann, J. A. (1994). Insect immunity. Septic injury of Drosophila induces the

synthesis of a potent antifungal peptide with sequence homology to plant antifungal peptides. J Biol Chem *269*, 33159-33163.

Ferrandon, D., Imler, J. L., and Hoffmann, J. A. (2004). Sensing infection in Drosophila: Toll and beyond. Semin Immunol *16*, 43-53.

Ferrandon, D., Jung, A. C., Criqui, M., Lemaitre, B., Uttenweiler-Joseph, S., Michaut, L., Reichhart, J., and Hoffmann, J. A. (1998). A drosomycin-GFP reporter transgene reveals a local immune response in Drosophila that is not dependent on the Toll pathway. Embo J *17*, 1217-1227.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res *8*, 967-974.

Franke-Fayard, B., Trueman, H., Ramesar, J., Mendoza, J., van der Keur, M., van der Linden, R., Sinden, R. E., Waters, A. P., and Janse, C. J. (2004). A Plasmodium berghei reference line that constitutively expresses GFP at a high level throughout the complete life cycle. Mol Biochem Parasitol *137*, 23-33.

Frevert, U. (2004). Sneaking in through the back entrance: the biology of malaria liver stages. Trends Parasitol *20*, 417-424.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., *et al.* (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature *419*, 498-511.

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet *29*, 482-486.

Ghosh, A., Edwards, M. J., and Jacobs-Lorena, M. (2000). The journey of the malaria parasite in the mosquito: hopes for the new century. Parasitol Today *16*, 196-201.

Girardin, S. E., Boneca, I. G., Viala, J., Chamaillard, M., Labigne, A., Thomas, G., Philpott, D. J., and Sansonetti, P. J. (2003). Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. J Biol Chem *278*, 8869-8872.

Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. Plant Physiol *124*, 1570-1581.

Goltsev, Y., Hsiong, W., Lanzaro, G., and Levine, M. (2004). Different combinations of gap repressors for common stripes in Anopheles and Drosophila embryos. Dev Biol *275*, 435-446.

Gomez, S. M., Eiglmeier, K., Segurens, B., Dehoux, P., Couloux, A., Scarpelli, C., Wincker, P., Weissenbach, J., Brey, P. T., and Roth, C. W. (2005). Pilot Anopheles gambiae full-length cDNA study: sequencing and initial characterization of 35,575 clones. Genome Biol *6*, R39.

Goodisman, M. A., Isoe, J., Wheeler, D. E., and Wells, M. A. (2005). Evolution of insect metamorphosis: a microarray-based study of larval and adult gene expression in the ant Camponotus festinatus. Evolution Int J Org Evolution *59*, 858-870.

Gottar, M., Gobert, V., Michel, T., Belvin, M., Duyk, G., Hoffmann, J. A., Ferrandon, D., and Royet, J. (2002). The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. Nature *416*, 640-644.

Greenwood, B., and Mutabingwa, T. (2002). Malaria in 2002. Nature *415*, 670-672.

Grossman, G. L., Rafferty, C. S., Clayton, J. R., Stevens, T. K., Mukabayire, O., and Benedict, M. Q. (2001). Germline transformation of the malaria vector, Anopheles gambiae, with the piggyBac transposable element. Insect Mol Biol *10*, 597-604.

Han, Y. S., Thompson, J., Kafatos, F. C., and Barillas-Mury, C. (2000). Molecular interactions between Anopheles stephensi midgut cells and Plasmodium berghei: the time bomb theory of ookinete invasion of mosquitoes. Embo J *19*, 6030-6040.

Hashimoto, C., Hudson, K. L., and Anderson, K. V. (1988). The Toll gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. Cell *52*, 269-279.

Hemingway, J., Hawkes, N. J., McCarroll, L., and Ranson, H. (2004). The molecular basis of insecticide resistance in mosquitoes. Insect Biochem Mol Biol *34*, 653-665.

Hemingway, J., and Ranson, H. (2000). Insecticide resistance in insect vectors of human disease. Annu Rev Entomol *45*, 371-391.

Hetru, C., Troxler, L., and Hoffmann, J. A. (2003). Drosophila melanogaster antimicrobial defense. J Infect Dis *187 Suppl 2*, S327-334.

Hild, M., Beckmann, B., Haas, S. A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., *et al.* (2003). An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. Genome Biol *5*, R3.

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. (2000). Genomic analysis of gene expression in C. elegans. Science *290*, 809-812.

Hoffman, S. L., Subramanian, G. M., Collins, F. H., and Venter, J. C. (2002). Plasmodium, human and Anopheles genomics and malaria. Nature *415*, 702-709.

Hoffmann, J. A. (2003). The immune response of Drosophila. Nature *426*, 33-38.

Hoffmann, J. A., and Reichhart, J. M. (2002). Drosophila innate immunity: an evolutionary perspective. Nat Immunol *3*, 121-126.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., *et al.* (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129-149.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., *et al.* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol *19*, 342-347.

Huizinga, E. G., Tsuji, S., Romijn, R. A., Schiphorst, M. E., de Groot, P. G., Sixma, J. J., and Gros, P. (2002). Structures of glycoprotein Ibalpha and its complex with von Willebrand factor A1 domain. Science *297*, 1176-1179.

Hultmark, D. (2003). Drosophila immunity: paths and patterns. Curr Opin Immunol *15*, 12-19.

Inohara, N., Ogura, Y., Fontalba, A., Gutierrez, O., Pons, F., Crespo, J., Fukase, K., Inamura, S., Kusumoto, S., Hashimoto, M., *et al.* (2003). Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. J Biol Chem *278*, 5509-5512.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., *et al.* (2005). Multiple-laboratory comparison of microarray platforms. Nat Methods *2*, 345-350.

Iwanaga, S., and Lee, B. L. (2005). Recent advances in the innate immunity of invertebrate animals. J Biochem Mol Biol *38*, 128-150.

Janeway, C. A., Jr., and Medzhitov, R. (2002). Innate immune recognition. Annu Rev Immunol *20*, 197-216.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. Genome Res *12*, 37-46.

Jiao, Y., Ma, L., Strickland, E., and Deng, X. W. (2005). Conservation and Divergence of Light-Regulated Genome Expression Patterns during Seedling Development in Rice and Arabidopsis. Plant Cell *17*, 3239-3256.

Kajava, A. V. (1998). Structural diversity of leucine-rich repeat proteins. J Mol Biol *277*, 519-527.

Kang, D., Romans, P., and Lee, J. Y. (1996). Analysis of a lysozyme gene from the malaria vector mosquito, Anopheles gambiae. Gene *174*, 239-244.

Kappe, S. H., Kaiser, K., and Matuschewski, K. (2003). The Plasmodium sporozoite journey: a rite of passage. Trends Parasitol *19*, 135-143.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Khusmith, S., Sedegah, M., and Hoffman, S. L. (1994). Complete protection against Plasmodium yoelii by adoptive transfer of a CD8+ cytotoxic T-cell clone recognizing sporozoite surface protein 2. Infect Immun *62*, 2979-2983.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. Am J Trop Med Hyg *70*, 486-498.

Knell, A. J., and Wellcome Tropical Institute. (1991). Malaria : a publication of the tropical programme of the Wellcome Trust (Oxford ; New York: Oxford University Press).

Kobe, B., and Deisenhofer, J. (1993). Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Nature *366*, 751-756.

Kobe, B., and Deisenhofer, J. (1994). The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci *19*, 415-421.

Kobe, B., and Deisenhofer, J. (1995). Proteins with leucine-rich repeats. Curr Opin Struct Biol *5*, 409-416.

Kobe, B., and Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol *11*, 725-732.

Kocks, C., Cho, J. H., Nehme, N., Ulvila, J., Pearson, A. M., Meister, M., Strom, C., Conto, S. L., Hetru, C., Stuart, L. M., *et al.* (2005). Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in Drosophila. Cell *123*, 335-346.

Koutsos, A. (2002) Identification of the LTR retrotransposons in the genome of Anopheles gambiae & Design of AnoBase, the new relational database of the Anopheles species, Master's thesis, University of Crete, Heraklion.

Kriventseva, E. V., Koutsos, A. C., Blass, C., Kafatos, F. C., Christophides, G. K., and Zdobnov, E. M. (2005). AnoEST: toward A. gambiae functional genomics. Genome Res *15*, 893-899.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., *et al.* (2004). The EMBL Nucleotide Sequence Database. Nucleic Acids Res *32*, D27-30.

Kumar, S., Christophides, G. K., Cantera, R., Charles, B., Han, Y. S., Meister, S., Dimopoulos, G., Kafatos, F. C., and Barillas-Mury, C. (2003). The role of reactive oxygen species on Plasmodium melanotic encapsulation in Anopheles gambiae. Proc Natl Acad Sci U S A *100*, 14139-14144.

Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics *18*, 405-412.

Kylsten, P., Samakovlis, C., and Hultmark, D. (1990). The cecropin locus in Drosophila; a compact gene cluster involved in the response to infection. Embo J *9*, 217-224.

Langer, R. C., and Vinetz, J. M. (2001). Plasmodium ookinete-secreted chitinase and parasite penetration of the mosquito peritrophic matrix. Trends Parasitol *17*, 269-272.
Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. Nat Methods *2*, 337-344.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res *14*, 1085-1094.

Lee, S. Y., Wang, R., and Soderhall, K. (2000). A lipopolysaccharide- and beta-1,3-glucan-binding protein from hemocytes of the freshwater crayfish Pacifastacus leniusculus. Purification, characterization, and cDNA cloning. J Biol Chem *275*, 1337-1343.

Lemaitre, B., Kromer-Metzger, E., Michaut, L., Nicolas, E., Meister, M., Georgel, P., Reichhart, J. M., and Hoffmann, J. A. (1995). A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the Drosophila host defense. Proc Natl Acad Sci U S A *92*, 9465-9469.

Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M., and Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. Cell *86*, 973-983.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. (2006). SMART 5: domains in the context of genomes and networks. Nucleic Acids Res *34*, D257-260.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. Nucleic Acids Res *32*, D142-144.

Levashina, E. A., Moita, L. F., Blandin, S., Vriend, G., Lagueux, M., and Kafatos, F. C. (2001). Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. Cell *104*, 709-718.

Levashina, E. A., Ohresser, S., Bulet, P., Reichhart, J. M., Hetru, C., and Hoffmann, J. A. (1995). Metchnikowin, a novel immune-inducible proline-rich peptide from Drosophila with antibacterial and antifungal properties. Eur J Biochem *233*, 694-700.

Li, D., Scherfer, C., Korayem, A. M., Zhao, Z., Schmidt, O., and Theopold, U. (2002). Insect hemolymph clotting: evidence for interaction between the coagulation system and the prophenoloxidase activating cascade. Insect Biochem Mol Biol *32*, 919-928.

Ligoxygakis, P., Pelte, N., Ji, C., Leclerc, V., Duvic, B., Belvin, M., Jiang, H., Hoffmann, J. A., and Reichhart, J. M. (2002). A serpin mutant links Toll activation to melanization in the host defence of Drosophila. Embo J *21*, 6330-6337.

Luna, C., Wang, X., Huang, Y., Zhang, J., and Zheng, L. (2002). Characterization of four Toll related genes during development and immune responses in Anopheles gambiae. Insect Biochem Mol Biol *32*, 1171-1179.

Ma, C., and Kanost, M. R. (2000). A beta1,3-glucan recognition protein from an insect, Manduca sexta, agglutinates microorganisms and activates the phenoloxidase cascade. J Biol Chem *275*, 7505-7514.

Ma, L., Chen, C., Liu, X., Jiao, Y., Su, N., Li, L., Wang, X., Cao, M., Sun, N., Zhang, X., *et al.* (2005). A microarray analysis of the rice transcriptome and its comparison to Arabidopsis. Genome Res *15*, 1274-1283.

Marino, M., Braun, L., Cossart, P., and Ghosh, P. (1999). Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes. Mol Cell *4*, 1063-1072.

Marinotti, O., Calvo, E., Nguyen, Q. K., Dissanayake, S., Ribeiro, J. M., and James, A. A. (2006). Genome-wide analysis of gene expression in adult Anopheles gambiae. Insect Mol Biol *15*, 1-12.

Marinotti, O., Nguyen, Q. K., Calvo, E., James, A. A., and Ribeiro, J. M. (2005). Microarray analysis of genes showing variable expression following a blood meal in Anopheles gambiae. Insect Mol Biol *14*, 365-373.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I., and Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet *36*, 197-204.

McGuinness, D. H., Dehal, P. K., and Pleass, R. J. (2003). Pattern recognition molecules and innate immunity to parasites. Trends Parasitol *19*, 312-319.

Meis, J. F., Wismans, P. G., Jap, P. H., Lensen, A. H., and Ponnudurai, T. (1992). A scanning electron microscopic study of the sporogonic development of Plasmodium falciparum in Anopheles stephensi. Acta Trop *50*, 227-236.

Meister, M., and Lagueux, M. (2003). Drosophila blood cells. Cell Microbiol *5*, 573-580.

Meister, S., Kanzok, S. M., Zheng, X. L., Luna, C., Li, T. R., Hoa, N. T., Clayton, J. R., White, K. P., Kafatos, F. C., Christophides, G. K., and Zheng, L. (2005). Immune

signaling pathways regulating bacterial and malaria parasite infection of the mosquito Anopheles gambiae. Proc Natl Acad Sci U S A *102*, 11420-11425.

Meister, S., Koutsos, A. C., and Christophides, G. K. (2004). The Plasmodium parasite--a 'new' challenge for insect innate immunity. Int J Parasitol *34*, 1473-1482.

Michel, K., Budd, A., Pinto, S., Gibson, T. J., and Kafatos, F. C. (2005). Anopheles gambiae SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. EMBO Rep *6*, 891-897.

Michel, T., Reichhart, J. M., Hoffmann, J. A., and Royet, J. (2001). Drosophila Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein. Nature *414*, 756-759.

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. (2004). DDBJ in the stream of various biological data. Nucleic Acids Res *32*, D31-34.

Moita, L. F., Wang-Sattler, R., Michel, K., Zimmermann, T., Blandin, S., Levashina, E. A., and Kafatos, F. C. (2005). In vivo identification of novel regulators and conserved pathways of phagocytosis in A. gambiae. Immunity *23*, 65-73.

Mongin, E., Louis, C., Holt, R. A., Birney, E., and Collins, F. H. (2004). The Anopheles gambiae genome: an update. Trends Parasitol *20*, 49-52.

Monnerat, A. T., Machado, M. P., Vale, B. S., Soares, M. J., Lima, J. B., Lenzi, H. L., and Valle, D. (2002). Anopheles albitarsis embryogenesis: morphological identification of major events. Mem Inst Oswaldo Cruz *97*, 589-596.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., *et al.* (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res *31*, 315-318.

Nagai, T., and Kawabata, S. (2000). A link between blood coagulation and prophenol oxidase activation in arthropod host defense. J Biol Chem *275*, 29264-29267.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell *2*, 279-289.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng *10*, 1-6.

Nozais, J. P. (2003). The origin and dispersion of human parasitic diseases in the old world (Africa, Europe and Madagascar). Mem Inst Oswaldo Cruz *98 Suppl 1*, 13-19.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., *et al.* (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature *420*, 563-573.

Onfelt Tingvall, T., Roos, E., and Engstrom, Y. (2001). The imd gene is required for local Cecropin expression in Drosophila barrier epithelia. EMBO Rep *2*, 239-243.

Osta, M. A., Christophides, G. K., and Kafatos, F. C. (2004). Effects of mosquito genes on Plasmodium development. Science *303*, 2030-2032.

Pancer, Z., Amemiya, C. T., Ehrhardt, G. R., Ceitlin, J., Gartland, G. L., and Cooper, M. D. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature *430*, 174-180.

Park, P. J., Butte, A. J., and Kohane, I. S. (2002). Comparing expression profiles of genes with similar promoter regions. Bioinformatics *18*, 1576-1584.

Paskewitz, S., and Riehle, M. A. (1994). Response of Plasmodium refractory and susceptible strains of Anopheles gambiae to inoculated Sephadex beads. Dev Comp Immunol *18*, 369-375.

Paskewitz, S. M., Brown, M. R., Lea, A. O., and Collins, F. H. (1988). Ultrastructure of the encapsulation of Plasmodium cynomolgi (B strain) on the midgut of a refractory strain of Anopheles gambiae. J Parasitol *74*, 432-439.

Ramet, M., Manfruelli, P., Pearson, A., Mathey-Prevot, B., and Ezekowitz, R. A. (2002). Functional genomic analysis of phagocytosis and identification of a Drosophila receptor for E. coli. Nature *416*, 644-648.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol *314*, 1041-1052.

Ribeiro, J. M., Topalis, P., and Louis, C. (2004). AnoXcel: an Anopheles gambiae protein database. Insect Mol Biol *13*, 449-457.

Richman, A. M., Bulet, P., Hetru, C., Barillas-Mury, C., Hoffmann, J. A., and Kafalos, F. C. (1996). Inducible immune factors of the vector mosquito Anopheles gambiae: biochemical purification of a defensin antibacterial peptide and molecular cloning of preprodefensin cDNA. Insect Mol Biol *5*, 203-210.

Richman, A. M., Dimopoulos, G., Seeley, D., and Kafatos, F. C. (1997). Plasmodium activates the innate immune response of Anopheles gambiae mosquitoes. Embo J *16*, 6114-6119.

Ridley, R. G. (2002). Medical need, scientific opportunity and the drive for antimalarial drugs. Nature *415*, 686-693.

Riehle, M. M., Markianos, K., Niare, O., Xu, J., Li, J., Toure, A. M., Podiougou, B., Oduol, F., Diawara, S., Diallo, M., *et al.* (2006). Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. Science *312*, 577-579.

Sachs, J., and Malaney, P. (2002). The economic and social burden of malaria. Nature *415*, 680-685.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Sambrook, J., and Russell, D. W. (2001). Molecular cloning : a laboratory manual, 3rd edn (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press).

Schnare, M., Barton, G. M., Holt, A. C., Takeda, K., Akira, S., and Medzhitov, R. (2001). Toll-like receptors control activation of adaptive immune responses. Nat Immunol *2*, 947-950.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A *95*, 5857-5864.

Schwartz, A., and Koella, J. C. (2002). Melanization of plasmodium falciparum and C-25 sephadex beads by field-caught Anopheles gambiae (Diptera: Culicidae) from southern Tanzania. J Med Entomol *39*, 84-88.

Shahabuddin, M., and Pimenta, P. F. (1998). Plasmodium gallinaceum preferentially invades vesicular ATPase-expressing cells in Aedes aegypti midgut. Proc Natl Acad Sci U S A *95*, 3385-3389.

Shao, L., Devenport, M., and Jacobs-Lorena, M. (2001). The peritrophic matrix of hematophagous insects. Arch Insect Biochem Physiol *47*, 119-125.

Shen, Z., Edwards, M. J., and Jacobs-Lorena, M. (2000). A gut-specific serine protease from the malaria vector Aanopheles gambiae is downregulated after blood ingestion. Insect Mol Biol *9*, 223-229.

Shen, Z., and Jacobs-Lorena, M. (1997). Characterization of a novel gut-specific chitinase gene from the human malaria vector Anopheles gambiae. J Biol Chem *272*, 28895-28900.

Shen, Z., and Jacobs-Lorena, M. (1998). A type I peritrophic matrix protein from the malaria vector Anopheles gambiae binds to chitin. Cloning, expression, and characterization. J Biol Chem *273*, 17665-17670.

Silverman, N., Zhou, R., Stoven, S., Pandey, N., Hultmark, D., and Maniatis, T. (2000). A Drosophila IkappaB kinase complex required for Relish cleavage and antibacterial immunity. Genes Dev *14*, 2461-2471.

Sim, C., Hong, Y. S., Vanlandingham, D. L., Harker, B. W., Christophides, G. K., Kafatos, F. C., Higgs, S., and Collins, F. H. (2005). Modulation of Anopheles gambiae gene expression in response to o'nyong-nyong virus infection. Insect Mol Biol *14*, 475-481.

Sinden, R. E. (1999). Plasmodium differentiation in the mosquito. Parassitologia *41*, 139-148.

Sinden, R. E., Winger, L., Carter, E. H., Hartley, R. H., Tirawanchai, N., Davies, C. S., Moore, J., and Sluiters, J. F. (1987). Ookinete antigens of Plasmodium berghei: a light and electron-microscope immunogold study of expression of the 21 kDa determinant recognized by a transmission-blocking antibody. Proc R Soc Lond B Biol Sci *230*, 443-458.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Soderhall, K., and Cerenius, L. (1998). Role of the prophenoloxidase-activating system in invertebrate immunity. Curr Opin Immunol *10*, 23-28.

Stearman, R. S., Dwyer-Nield, L., Zerbe, L., Blaine, S. A., Chan, Z., Bunn, P. A., Jr., Johnson, G. L., Hirsch, F. R., Merrick, D. T., Franklin, W. A., *et al.* (2005). Analysis of Orthologous Gene Expression between Human Pulmonary Adenocarcinoma and a Carcinogen-Induced Murine Model. Am J Pathol *167*, 1763-1775.

Sterrenburg, E., Turk, R., Boer, J. M., van Ommen, G. B., and den Dunnen, J. T. (2002). A common reference for cDNA microarray hybridizations. Nucleic Acids Res *30*, e116.

Stoughton, R. B. (2005). Applications of DNA microarrays in biology. Annu Rev Biochem *74*, 53-82.

Stoven, S., Ando, I., Kadalayil, L., Engstrom, Y., and Hultmark, D. (2000). Activation of the Drosophila NF-kappaB factor Relish by rapid endoproteolytic cleavage. EMBO Rep *1*, 347-352.

Tahar, R., Boudin, C., Thiery, I., and Bourgouin, C. (2002). Immune response of Anopheles gambiae to the early sporogonic stages of the human malaria parasite Plasmodium falciparum. Embo J *21*, 6673-6680.

Takehana, A., Katsuyama, T., Yano, T., Oshima, Y., Takada, H., Aigaki, T., and Kurata, S. (2002). Overexpression of a pattern-recognition receptor, peptidoglycan-recognition protein-LE, activates imd/relish-mediated antibacterial defense and the prophenoloxidase cascade in Drosophila larvae. Proc Natl Acad Sci U S A *99*, 13705-13710.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907-2912.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res *31*, 5676-5684.

Team, R. D. C. (2006). R: A language and Environment for statistical computing.
Theopold, U., Li, D., Fabbri, M., Scherfer, C., and Schmidt, O. (2002). The coagulation of insect hemolymph. Cell Mol Life Sci *59*, 363-372.

Thomasova, D., Ton, L. Q., Copley, R. R., Zdobnov, E. M., Wang, X., Hong, Y. S., Sim, C., Bork, P., Kafatos, F. C., and Collins, F. H. (2002). Comparative genomic analysis in the region of a major Plasmodium-refractoriness locus of Anopheles gambiae. Proc Natl Acad Sci U S A *99*, 8179-8184.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., *et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol *23*, 137-144.

Topalis, P., Koutsos, A., Dialynas, E., Kiamos, C., Hope, L. K., Strode, C., Hemingway, J., and Louis, C. (2005). AnoBase: a genetic and biological database of anophelines. Insect Mol Biol *14*, 591-597.

Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. FEBS Lett *451*, 142-146.

Tzou, P., Ohresser, S., Ferrandon, D., Capovilla, M., Reichhart, J. M., Lemaitre, B., Hoffmann, J. A., and Imler, J. L. (2000). Tissue-specific inducible expression of antimicrobial peptide genes in Drosophila surface epithelia. Immunity *13*, 737-748.

van der Krol, A. R., Mur, L. A., Beld, M., Mol, J. N., and Stuitje, A. R. (1990). Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant Cell *2*, 291-299.

Vasselon, T., and Detmers, P. A. (2002). Toll receptors: a central element in innate immune responses. Infect Immun *70*, 1033-1041.

Vernick, K. D., Fujioka, H., Seeley, D. C., Tandler, B., Aikawa, M., and Miller, L. H. (1995). Plasmodium gallinaceum: a refractory mechanism of ookinete killing in the mosquito, Anopheles gambiae. Exp Parasitol *80*, 583-595.

Vesanto, J., and Alhoniemi, E. (2000). Clustering of the self-organising map. IEEE Transactions on Neural Networks *11*, 586-600.

Vizioli, J., Bulet, P., Hoffmann, J. A., Kafatos, F. C., Muller, H. M., and Dimopoulos, G. (2001a). Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector Anopheles gambiae. Proc Natl Acad Sci U S A *98*, 12630-12635.

Vizioli, J., Richman, A. M., Uttenweiler-Joseph, S., Blass, C., and Bulet, P. (2001b). The defensin peptide of the malaria vector mosquito Anopheles gambiae: antimicrobial activities and expression in adult mosquitoes. Insect Biochem Mol Biol *31*, 241-248.

Vlachou, D., Schlegelmilch, T., Christophides, G. K., and Kafatos, F. C. (2005). Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. Curr Biol *15*, 1185-1195.

Vlachou, D., Zimmermann, T., Cantera, R., Janse, C. J., Waters, A. P., and Kafatos, F. C. (2004). Real-time, in vivo analysis of malaria ookinete locomotion and mosquito midgut invasion. Cell Microbiol *6*, 671-685.

Volz, J., Mueller, H. M., Zdanowicz, A., Kafatos, F. C., and Osta, M. A. (2006). A genetic module regulates the malanization response of *Anopheles* to *Plasmodium*. Cellular Microbiology *in press*.

Volz, J., Osta, M. A., Kafatos, F. C., and Muller, H. M. (2005). The roles of two clip domain serine proteases in innate immune responses of the malaria vector Anopheles gambiae. J Biol Chem *280*, 40161-40168.

Vontas, J., Blass, C., Koutsos, A. C., David, J. P., Kafatos, F. C., Louis, C., Hemingway, J., Christophides, G. K., and Ranson, H. (2005). Gene expression in insecticide resistant and susceptible Anopheles gambiae strains constitutively or after insecticide exposure. Insect Mol Biol *14*, 509-521.

Waller, R. F., Keeling, P. J., Donald, R. G., Striepen, B., Handman, E., Lang-Unnasch, N., Cowman, A. F., Besra, G. S., Roos, D. S., and McFadden, G. I. (1998). Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum. Proc Natl Acad Sci U S A *95*, 12352-12357.

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet *5*, 276-287.

Wicker, C., Reichhart, J. M., Hoffmann, D., Hultmark, D., Samakovlis, C., and Hoffmann, J. A. (1990). Insect immunity. Characterization of a Drosophila cDNA encoding a novel member of the diptericin family of immune peptides. J Biol Chem *265*, 22493-22498.

Wirth, D. F. (2002). Biological revelations. Nature *419*, 495-496.

Xu, P. X., Zwiebel, L. J., and Smith, D. P. (2003). Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, Anopheles gambiae. Insect Mol Biol *12*, 549-560.

Yoshida, H., Kinoshita, K., and Ashida, M. (1996). Purification of a peptidoglycan recognition protein from hemolymph of the silkworm, Bombyx mori. J Biol Chem *271*, 13854-13860.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., *et al.* (2002). Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science *298*, 149-159.

Zheng, L., Cornel, A. J., Wang, R., Erfle, H., Voss, H., Ansorge, W., Kafatos, F. C., and Collins, F. H. (1997). Quantitative trait loci for refractoriness of Anopheles gambiae to Plasmodium cynomolgi B. Science *276*, 425-428.

Zheng, L., Wang, S., Romans, P., Zhao, H., Luna, C., and Benedict, M. Q. (2003). Quantitative trait loci in Anopheles gambiae controlling the encapsulation response against Plasmodium cynomolgi Ceylon. BMC Genet *4*, 16.

Zheng, L., Whang, L. H., Kumar, V., and Kafatos, F. C. (1995). Two genes encoding midgut-specific maltase-like polypeptides from Anopheles gambiae. Exp Parasitol *81*, 272-283.

Zheng, X. L., and Zheng, A. L. (2002). Genomic organization and regulation of three cecropin genes in Anopheles gambiae. Insect Mol Biol *11*, 517-525.

Zhu, Y., Wang, Y., Gorman, M. J., Jiang, H., and Kanost, M. R. (2003). Manduca sexta serpin-3 regulates prophenoloxidase activation in response to infection by inhibiting prophenoloxidase-activating proteinases. J Biol Chem *278*, 46556-46564.

Zieler, H., and Dvorak, J. A. (2000). Invasion in vitro of mosquito midgut cells by the malaria parasite proceeds by a conserved mechanism and results in death of the invaded midgut cells. Proc Natl Acad Sci U S A *97*, 11516-11521.

Supplementary DVD

The supplementary DVD contains the datasets with the TCLAG contig lists that are mentioned in chapter 3 of the thesis. Plase refer to the file readme.pdf for more information.

List of Publications

Koutsos, A. C., Blass, C, Meister, S., Schmidt, S., Soares, M. B., Collins, F. H., Benes, V., Zdobnov, E. M., Kafatos, F. C., Christophides, G. K. Lifecycle transcriptomics of the malaria mosquito *Anopheles gambiae* and comparison with the fruitfly *Drosophila melanogaster*. Manuscript in preparation.

Kriventseva, E. V.*, Koutsos, A. C.*, Blass, C., Kafatos, F. C., Christophides, G. K., and Zdobnov, E. M. (2005). AnoEST: toward A. gambiae functional genomics. Genome Res *15*, 893-899.

Vontas, J., Blass, C., Koutsos, A. C., David, J. P., Kafatos, F. C., Louis, C., Hemingway, J., Christophides, G. K., and Ranson, H. (2005). Gene expression in insecticide resistant and susceptible Anopheles gambiae strains constitutively or after insecticide exposure. Insect Mol Biol *14*, 509-521.

Meister, S.*, Koutsos, A. C.*, and Christophides, G. K. (2004). The Plasmodium parasite--a 'new' challenge for insect innate immunity. Int J Parasitol *34*, 1473-1482.

*equal contributing authors

AnoEST: Toward *A. gambiae* functional genomics

Evgenia V. Kriventseva,[1] Anastasios C. Koutsos,[1] Claudia Blass, Fotis C. Kafatos, George K. Christophides, and Evgeny M. Zdobnov[2]

European Molecular Biology Laboratory, D69117 Heidelberg, Germany

Here, we present an analysis of 215,634 EST and cDNA sequences of a major vector of human malaria *Anopheles gambiae* structured into the AnoEST database. The expressed sequences are grouped into clusters using genomic sequence as template and associated with inferred functional annotation, including the following: corresponding Ensembl gene prediction, putative orthologous genes in other species, homology to known proteins, protein domains, associated Gene Ontology terms, and corresponding classification into broad GO-slim functional groups. AnoEST is a vital resource for interpretation of expression profiles derived using recently developed *A. gambiae* cDNA microarrays. Using these cDNA microarrays, we have experimentally confirmed the expression of 7961 clusters during mosquito development. Of these, 3100 are not associated with currently predicted genes. Moreover, we found that clusters with confirmed expression are nonbiased with respect to the current gene annotation or homology to known proteins. Consequently, we expect that many as yet unconfirmed clusters are likely to be actual *A. gambiae* genes. [AnoEST is publicly available at http://komar.embl.de, and is also accessible as a Distributed Annotation Service (DAS).]

Blood-feeding anopheline mosquitoes are obligatory vectors for the transmission of the malaria parasites of the genus *Plasmodium*. The parasites undergo asexual development within mammalian hosts and produce gametocytes which, when ingested by the mosquito, initiate the sexual cycle that culminates with production of sporozoites. In turn, an infected mosquito takes another bloodmeal and sporozoites are released into the circulation of a naive host, thus completing the transmission cycle. Human malaria causes over 1 million deaths every year in the developing world. Recently, in recognition of the great importance of *Anopheles gambiae* in global health, its genome has been sequenced by an international scientific consortium (Holt et al. 2002), and transcriptomic approaches were initiated with the sequencing of Expressed Sequence Tags (ESTs) prepared from cultured cells (Dimopoulos et al. 2000). Four thousand ESTs were used to construct the first mosquito cDNA microarray, the 4K microarray platform (Dimopoulos et al. 2002). These arrays were used to detect genes that are up-regulated in the mosquito, specifically during infection with parasites and bacteria (Dimopoulos et al. 2002) and to identify differences between parasite-susceptible and refractory mosquitoes (Kumar et al. 2003). However, insufficient annotation of the EST sequences hindered such studies and greatly limited the capacity of researchers to derive appropriate interpretations. In the context of the *Anopheles* genome project, nearly 83,000 ESTs from naive and blood-fed adult mosquitoes were sequenced (Holt et al. 2002), and in silico analysis of these data detected genes up-regulated in the mosquitoes after a blood meal (Ribeiro et al. 2004). Furthermore, nearly 63,000 single reads from a full-length cDNA library were recently deposited in nucleotide databases by Genoscope (http://www.genoscope.org/). Two other EST libraries were constructed from pooled developmental stages of *A. gambiae* (NAP1) or adult heads (NAH), and clones from these libraries are currently being sequenced (G.K. Christophides, unpubl.; F. Collins, unpubl.). Twenty thousand of these ESTs were used to build a new cDNA microarray platform (20K or MMC1), which is currently used in various experimental approaches to identify genes that are temporally and spatially regulated in mosquitoes during development, parasite and viral infection, and insecticide treatment (G.K. Christophides, unpubl.). The increasing amount of information obtained from such studies necessitated the development of computational approaches to provide functional annotation and interpretation of the derived data.

Here, we report a large-scale study of malaria mosquito *A. gambiae* EST and cDNA sequences structured into the newly developed AnoEST database. Using these cDNA microarray data in conjunction with AnoEST, we have experimentally confirmed expression of 7961 clusters during mosquito development. Of these, 3100 are not associated with currently predicted genes (Holt et al. 2002; Birney et al. 2004). Moreover, we found that clusters with confirmed expression are nonbiased with respect to the current gene annotation or homology to known proteins, and consequently, we might expect that many of the unconfirmed clusters are likely to be actual *A. gambiae* genes. The AnoEST resource is a vital resource for the interpretation of expression profiles derived using the *A. gambiae* cDNA microarrays, providing inferred functional annotation of the expressed genomic loci, including similarities to known proteins, protein domains, and Gene Ontology (GO) (Ashburner et al. 2000) functional categories.

Results and Discussion

A. gambiae EST classification

We collected from public sequence databases (Benson et al. 2004; Kulikova et al. 2004; Miyazaki et al. 2004) 215,634 *A. gambiae* expressed sequences (178,618 from 5′-sequences and 37,015 from 3′-sequences) originating from 179,955 clones. Of these sequences, 211,468 were aligned to 593,349 regions on the nuclear or mitochondrial genome. For 203,812 expressed sequences, a unique genomic origin could be recognized. We clus-

[1]These authors contributed equally to this work.
[2]Corresponding author.
E-mail zdobnov@embl.de; fax 49-6221-387-517.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3756405. Article published online ahead of print in May 2005. Freely available online through the *Genome Research* Immediate Open Access option.

Insect Molecular Biology (2005) **14**(5), 509–521

Gene expression in insecticide resistant and susceptible *Anopheles gambiae* strains constitutively or after insecticide exposure

J. Vontas,*‡ C. Blass,† A. C. Koutsos,† J.-P. David,‡ F. C. Kafatos,† C. Louis,*§ J. Hemingway‡ G. K. Christophides† and H. Ranson‡‡

**Institute of Molecular Biology and Biotechnology (IMBB-FORTH), Vassilika Vouton, Heraklion, Crete, Greece;* †*European Molecular Biology Laboratory, Heidelberg, Germany;* ‡*Vector Research, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK; and* §*Department of Biology, University of Crete, Heraklion, Crete, Greece*

Abstract

A microarray containing approximately 20 000 expressed sequence tags (ESTs; 11 760 unique EST clusters) from the malaria vector, *Anopheles gambiae*, was used to monitor differences in global gene expression in two insecticide resistant and one susceptible strains. Statistical analysis identified 77 ESTs that were differentially transcribed among the three strains. These include the cytochrome P450 *CYP314A1*, over-transcribed in the DDT resistant *ZAN/U* strain, and many genes that belong to families not usually associated with insecticide resistance, such as peptidases, sodium/calcium exchangers and genes implicated in lipid and carbohydrate metabolism. Short-term (6 and 10 h) effects of exposure of the pyrethroid resistant *RSP* strain to permethrin were also detected. Several genes belonging to enzyme families already implicated in insecticide or xenobiotic detoxification were induced, including the carboxylesterase *COEAE2F* gene and members of the UDP-glucuronosyl transferase and nitrilase families.

Keywords: *Anopheles gambiae*, insecticide resistance, microarray, detoxification.

doi: 10.1111/j.1365-2583.2005.00582.x

Received 8 February 2005; accepted after revision 27 April 2005. Correspondence: Hilary Ranson, Vector Research, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L35QA, UK. Tel: +44 151 705 3310; fax: +44 151 705 3369; email: Hranson@liverpool.ac.uk and John Vontas, Laboratory of Pesticide Science, Agricultural University of Athens, Iera Odos 75, Athens 11855, Greece. Tel: +30 21 05294546; fax: +30 21 05294514; e-mail: vontas@aua.gr

Introduction

Insecticides form an integral part of all national malaria control programs. Members of four insecticide classes are registered for use in indoor residual house spraying and the use of pyrethroid-impregnated bednets is actively promoted in many malaria endemic countries (WHO, 2000). Dichlorodiphenyltrichloroethane (DDT), while internationally banned for general use by the Persistent Organic Pollutants Treaty, is still registered for indoor house spraying and is regarded as an essential insecticide for current malaria control operations. The emergence and spread of insecticide resistance can have a devastating effect on these control measures. DDT resistance is widespread in many *Anopheles* species and in Africa, where 90% of malaria cases occur, three independent foci of pyrethroid resistance have emerged (Elissa *et al*., 1993; Vulule *et al*., 1994; Hargreaves *et al*., 2000).

Most studies on the molecular basis of insecticide resistance focus on two major mechanisms, which are changes in the sensitivity of insecticide targets in the nervous system and increases in the rate of insecticide detoxification. Both of these resistance mechanisms are clearly operating in *A. gambiae*. Mutations in a major insecticide target, the voltage gated sodium channel, confer resistance to DDT and pyrethroids in *A. gambiae* (Martinez Torres *et al*., 1998; Ranson *et al*., 2000a), and an amino acid substitution in acetylcholinesterase, the target site of carbamate and organophosphate insecticides, has been identified in West African populations (Weill *et al*., 2003). Metabolic resistance to insecticides in *A. gambiae* can be conferred by elevation in the activity of detoxifying enzymes, such as the glutathione transferase GSTE2 (Ranson *et al*., 2001) or members of the cytochrome P450 family (Nikou *et al*., 2003).

Many studies have suggested that the insecticide resistance phenotype evolves rapidly based on the selection of major effect genes (Daborn *et al*., 2002; ffrench-Constant *et al*., 2004). However recent genome-wide transcription profiling indicated that a broader range of genes may be involved and that insecticide resistance may be more complex than previously considered (Pedra *et al*., 2004). Genetic mapping has identified four major quantitative trait loci (QTL) associated with resistance to pyrethroids or DDT

Available online at www.sciencedirect.com
SCIENCE DIRECT
International Journal for Parasitology 34 (2004) 1473–1482
www.parasitology-online.com

Invited review

The *Plasmodium* parasite—a 'new' challenge for insect innate immunity

S. Meister[1], A.C. Koutsos[1], G.K. Christophides*

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received 29 April 2004; received in revised form 17 September 2004; accepted 1 October 2004

Abstract

Though lacking adaptive immunity, insects possess a powerful innate immune system, a genome-encoded defence machinery used to confront infections. Studies in the fruit fly *Drosophila melanogaster* revealed a remarkable capacity of the innate immune system to differentiate between and subsequently respond to different bacteria and fungi. However, hematophagous compared to non-hematophagous insects encounter additional blood-borne infectious agents, such as parasites and viruses, during their lifetime. *Anopheles* mosquitoes become infected with the malaria parasite *Plasmodium* during feeding on infected human hosts and may then transmit the parasite to new hosts during subsequent bites. Whether *Anopheles* has developed mechanisms to confront these infections is the subject of this review. Initially, we review our current understanding of innate immune reactions and give an overview of the *Anopheles* immune system as revealed through comparative genomic analyses. Then, we examine and discuss the capacity of mosquitoes to recognize and respond to infections, especially to *Plasmodium*, and finally, we explore approaches to investigate and potentially utilize the vector immune competence to prevent pathogen transmission. Such approaches constitute a new challenge for insect immunity research, a challenge for global health.

Keywords: *Anopheles*; *Plasmodium*; Malaria; Innate immunity; *LRIM1*; *CTL4*

1. Introduction

During their lifetime, metazoans encounter various microbial invaders, pathogenic or not. To keep them at bay, they developed an array of defence mechanisms, collectively referred to as the immune system. The evolutionary ancestral innate immune system is the organism's inherited primary line of defence, and as such, it depends on a given number of immune molecules. A different strategy of defence, the adaptive or acquired immune system, was discovered later in evolutionary history, and is based on the virtually unlimited resource of natural recombinant proteins (antibodies) created through genomic rearrangements. A primary difference between the two immune systems is that memory or specialization to cope with the variety of infective agents is obtained solely through evolution in the case of innate immunity, whereas adaptive immunity employs the additional strategies of clonal cell selection and proliferation to fight new invaders. It must be emphasized that apart from its own essential role in defence, innate immunity serves to trigger and direct the adaptive immune responses, as well as to gain time for the adaptive immune system to unfold its full effectiveness (Fearon and Locksley, 1996; Schnare et al., 2001).

Innate immunity relies on a limited number of receptors recognizing invariant molecules that either are on the surface of or are shed by microorganisms and are referred to as pathogen associated molecular patterns (PAMPs) (Janeway and Medzhitov, 2002). Here, we adopt this term, though not totally appropriate as similar PAMPs may be shared between both pathogenic and non-pathogenic microorganisms. Microbial PAMPs include molecules such as peptidoglycans (PGN), lipopolysaccharides (LPS) and β-1,3-glucans, as well as GPI (glycosyl phospatidyl inositol) anchors of protozoan parasites (Teixeira et al., 2002).

There are two types of innate immune responses: the humoral response with the antimicrobial peptides (AMPs) as its hallmark, and the cellular response that includes phagocytosis or encapsulation of the intruders. Thus, innate

* Corresponding author. Tel.: +49 6221 387440; fax: +49 6221 387306.
E-mail address: christop@embl.de (G.K. Christophides).
[1] These authors contributed equally to this review.

0020-7519/$30.00
doi:10.1016/j.ijpara.2004.10.004

Credits

Chapter 2 has been a collaborative research effort with Evgenia Kriventseva, Claudia Blass, Evgeny Zdobnov, George Christophides and Fotis Kafatos. Chapter 3 has been a collaborative research effort with Claudia Blass, Stephan Meister, Sabine Schmidt, Marcello Soares, Frank Collins, Vladimir Benes, Evgeny Zdonbnov, Fotis Kafatos and George Christophides. Chapter 4 contains information from a previously published study (Osta et al. 2004). Chapter 4 discussion contains information from *P. falciparum* studies (Cohuet, A et al, submitted for publication), for SRPN2 gene (Michel, K., unpublished observations), as well as results for LRIM1 and TEP1 from a collaboration with Stephanie Blandin and Elena Levashina (IBMC, Strassbourg).

Some pictures have been used from other studies. The copyrights of these images lie in their respective owners.