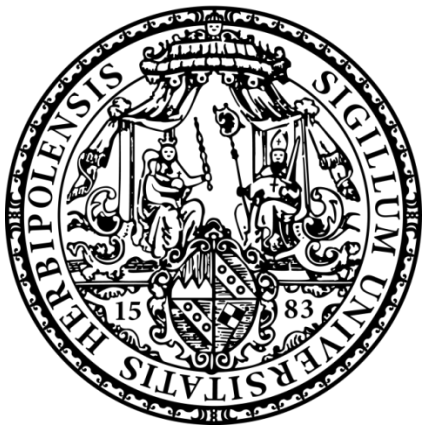
Re-annotation of *Camponotus floridanus* Genome and Characterization of Innate Immunity Transcriptome Responses to Bacterial Infections



DISSERTATION

zur Erlangung des naturwissenschaftlichen Doktorgrades

der Fakultät für Biologie,

Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von

Shishir Kumar Gupta

aus Lakhnau, Indien

Würzburg, Germany 2016

Eingereicht am:
.....

Mitglieder der Promotionskommission:

Vorsitzender:

Erster Gutachter: Prof. Dr. Thomas Dandekar

Zweiter Gutachter: Prof. Dr. Roy Gross

Tag des Promotionskolloquiums:

Doktorurkunde ausgehändigt am:

Erklärung (gem. § 4 Abs. 3 S. 3, 5 und 8 der Promotionsordnung)

Hiermit erkläre ich an Eides statt, die Dissertation “Re-annotation of *Camponotus floridanus* genome and characterization of innate immunity transcriptome responses to bacterial infections” eigenständig, d.h. insbesondere selbstständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben. Ich erkläre außerdem, dass dir Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat. Zusätzlich habe oder werde ich nicht versuchen neben diesem Abschluss einen weiteren Abschluss oder Qualifikation mit dieser Doktorarbeit zu erweben.

Ort, Datum

Würzburg, ……. Shishir Kumar Gupta

Dream is not that which you see while sleeping, it is something that does not let you sleep.

- Dr. A.P.J Abdul Kalam

Dedicated to my teacher Prof. Abhilash PC and my beloved parents.

Acknowledgements

Foremost, I owe my deepest gratitude to my mentor, Prof. Dr. Thomas Dandekar for his earnest guidance, and exposing me to this fascinating topic, constant attention, insightful comments, patience, personal encouragement and support throughout the course of this investigation and preparation of this thesis. His explicit style of functioning inspired me to work with utmost pleasure and freedom to explore on my own.

I bid my sincere thanks to Prof. Dr. Roy Gross for the valuable suggestions, discussions, critical comments, shaping the manuscripts and providing the experimental data which made the foundation of this thesis.

Besides my advisors, I am also grateful to Prof. Abhilash PC (Banaras Hindu University, India) and Dr. Shailendra K. Gupta (University of Rostock, Germany) for constant moral support, motivation and encouragement.

No word could do justice to thanks my colleague and wife Dr. Mugdha Srivastava for her love, fidelity, understanding and unconditional support during special difficult moments throughout this research.

Special thanks to my friend Dr. Bashir A. Akhoon for the long scientific and non-scientific discussions, constant motivation to each other, helping each other and working together since last eight years.

I am thankful to Dr. Chunguang Liang for his expert support for anything related to complex programming, guiding me for the cluster jobs and also helping me to deal with the non-work related issues.

I am thankful to Stefan Obermeier for his constant IT-support despite his health issues.

I am thankful to my colleagues and friends Dr. Naseem Muhammad, Dr. Edita Sarukhanyan, Dr. Zeeshan Ahmed, Christian Remmele, Johannes Balkenhol, Christian Luther and Dominik Schaack for their immense help, fun and the scientific and non-scientific discussions I used to have with them on various topics that ensued a relaxing time from research.

My heartful thanks to my friend Srikant Awasthi for his fidelity and visiting my parents during the time whenever it was needed.

I am thankful to Dr. Elena Bencúrová for her support, discussions and proofreading of this thesis.

Special thanks to Mercedes Hof for translating the summary into German.

Special thanks to Eva-Maria Fischer and Petra Narrog for their consistent support with all the administrative stuff and helping me to fill out hundreds of forms, their management skills made my life in Germany bit easier.

I gratefully acknowledge the financial support from DFG (GR 1243/8-1); and financial support from IZKF/B-243.

Thanks to BioMed Central, for licencing our article published in BMC Genomics under Creative Commons Attribution License 4.0 (CC BY 4.0).

I express my deep sense of thanks to the collaborators and the co-authors of the manuscripts, Prof. Dr. Heike Feldhaar, Prof. Dr. Andreas Vilcinskas, Dr. Frank Förster, Dr. Carolin Ratzka, and Maria Kupper, for their timely help and co-cooperation rendered to me at different stages of this work.

I take this opportunity to put across my indebtness and gratitude to my teachers and mentors Prof. Dr. Niels Grabe, Prof. Alok Dhawan, Dr. Nandita Singh, Prof. Dr. Mario Stanke, Dr. Paul Dijkwel, Dr. Julio Augusto Freyre-González, and Dr. Manish Kumar Gupta, for their role in my research career, advice, wishes and whole hearted support.

I wish to thank my all the friends especially Dr. Ashutosh K. Pandey, Ashwani Srivastava, Ashish Singh, Pramod K. Sharma, Deepak Yadav, Archana Singh, Merlin James, Neha Gottlieb, Nataraj Kalleda, Ayan Dhara, Dr. Pramod K. Verma, Sanjay Gupta, Dr. Farhad Guliyev, Alex and Thomas for their friendship, support, special care and warmth.

I would like to express my heart-felt gratitude to my family. In fact, none of this would have been possible without the constant source of love, concern, support and patience of my family. A very special thanks to my beloved mother for spiritually supporting me throughout my life and to my father, without whose unconditional love and encouragement I would not have considered a graduate career in life sciences research. Their wishes and prayers given the courage to me do research more effectually.

A very special thanks to my little baby 'Mishika Gupta', who has been the light of my life for the last one year and has given me the extra strength and motivation to get things done. Thanks for her calmness, to let me work and write at home.

The last word of thanks goes to 'The Almighty' for guiding me on the right path and providing me the basic sense and knowledge that helped me to successfully complete this thesis.

Summary

The sequencing of several ant genomes within the last six years open new research avenues for understanding not only the genetic basis of social species but also the complex systems such as immune responses in general. Similar to other social insects, ants live in cooperative colonies, often in high densities and with genetically identical or closely related individuals. The contact behaviours and crowd living conditions allow the disease to spread rapidly through colonies. Nevertheless, ants can efficiently combat infections by using diverse and effective immune mechanisms. However, the components of the immune system of carpenter ant *Camponotus floridanus* and also the factors in bacteria that facilitate infection are not well understood.

To form a better view of the immune repository and study the *C. floridanus* immune responses against the bacteria, experimental data from Illumina sequencing and mass-spectrometry (MS) data of haemolymph in normal and infectious conditions were analysed and integrated with the several bioinformatics approaches. Briefly, the tasks were accomplished in three levels. First, the *C. floridanus* genome was re-annotated for the improvement of the existing annotation using the computational methods and transcriptomics data. Using the homology based methods, the extensive survey of literature, and mRNA expression profiles, the immune repository of *C. floridanus* were established. Second, large-scale protein-protein interactions (PPIs) and signalling network of *C. floridanus* were reconstructed and analysed and further the infection induced functional modules in the networks were detected by mapping of the expression data over the networks. In addition, the interactions of the immune components with the bacteria were identified by reconstructing inter-species PPIs networks and the interactions were validated by literature. Third, the stage-specific MS data of larvae and worker ants were analysed and the differences in the immune response were reported.

Concisely, all the three omics levels resulted to multiple findings, for instance, re-annotation and transcriptome profiling resulted in the overall improvement of structural and functional annotation and detection of alternative splicing events, network analysis revealed the differentially expressed topologically important proteins and the active functional modules, MS data analysis revealed the stage specific differences in *C. floridanus* immune responses against bacterial pathogens.

Taken together, starting from re-annotation of *C. floridanus* genome, this thesis provides a transcriptome and proteome level characterization of ant *C. floridanus*, particularly focusing on the immune system responses to pathogenic bacteria from a biological and a bioinformatics point of view. This work can serve as a model for the integration of omics data focusing on the immuno-transcriptome of insects.

Zusammenfassung

Das Sequenzieren mehrerer Ameisen Genome innerhalb der letzten 6 Jahre eröffnete neue Forschungswege, um nicht nur die genetische Grundlade sozialer Arten, sondern auch komplexere Systeme wie generelle Immunantworten zu untersuchen. Ähnlich zu anderen sozialen Insekten leben Ameisen in Kolonien, oft mit einer sehr hohen Dichte mit genetisch übereinstimmenden oder nah verwandten Individuen. Das Sozialverhalten und die engen Lebensumstände führen dazu, dass sich Krankheiten in Kolonien schnell ausbreiten können. Dennoch können Ameisen mit der Nutzung ihrer komplexen Immunsystemmechanismen Infektionen effektiv abwehren. Die Zusammensetzung des Immunsystems der Rossameise *Camponotus floridanus* (*C. floridanus*) und die Faktoren der Bakterien, welche die Infektionen verursachen sind noch nicht gut untersucht.

Um einen besseren Überblick über die verschiedenen Gruppen der Immun- Gene zu bekommen und um die Immunantworten von *C. floridanus* gegen Bakterien zu untersuchen haben wir experimentelle Daten der Illumina Sequenzierung und der Massenspektrometrie (MS) aus der Hämolymphe unter normalen und unter infizierten Bedingungen analysiert und über verschiedene bioinformatische Ansätzen zusammengefasst. Die Aufgabe wurde in drei Ebenen unterteilt. Zuerst wurde das Genom von *C. floridanus* neu annotiert, die Verbesserung der existierenden Annotation wurde rechnerisch und mit Transkriptom- Daten erreicht. Mit der Nutzung der auf Homologie- basierenden Methoden, der umfassenden Überprüfung der Literatur und der Nutzung von mRNA Genexpressionsanalysen wurde für *C. floridanus* dieser Überblick erstellt. Anschließend wurden größere Protein- Protein- Interaktionen (PPI) und Signalnetzwerke von *C. floridanus* rekonstruiert und analysiert und daraufhin wurden die Infektions-induzierten funktionalen Module im Netzwerk entdeckt und die Expressionsdaten über Netzwerke abgebildet. Zusätzlich wurden die Anteile der Immunantwort bei der Interaktion mit Bakterien mittels der Rekonstruktion von zwischenartlichen PPI Netzwerken identifiziert und diese Interaktionen wurden mit Literaturwerten validiert. In der dritten und letzten Phase wurden Daten der Stadium- spezifischen Massenspektrometrie (MS) von Larven- und Arbeiterameisen analysiert und die Unterschiede in den Immunantworten aufgezeichnet.

Zusammengefasst lieferten alle drei Omiks- Ebenen jeweils viele Ergebnisse, zum Beispiel führte die neue Annotation und das Transkription- Profil zu einer generellen Verbesserung der strukturellen und funktionalen Annotation und dem Aufspüren von

alternativen Splicing- Ereignissen. Die Netzwerkanalyse deckte die unterschiedlich exprimierten topologisch wichtigen Proteine und die aktiven funktionalen Module auf, die Analyse der MS- Daten erbrachte Ergebnisse über die Stadium- spezifischen Unterschiede in der Immunantwort von *C. floridanus* gegen bakterielle Pathogene.

Rundum, beginnend mit der neuen Annotation des Genoms von *C. floridanus* stellt diese Arbeit eine Transkriptom- und Protein Charakterisierung der Ameise *C. floridanus* dar. Besonders lag der Fokus auf die Antworten des Immunsystems auf Pathogene Bakterien aus biologischer- und bioinformatischer Sicht. Diese Arbeit kann als Vorlage für die Integration von Omiks Daten dienen, welche sich auf die Immun- Transkriptome von Insekten fokussieren.

Contents

1

Introduction

1.1 Genome sequencing

Over the past decade and a half, the emergence of several Next-Generation Sequencing (NGS) techniques has transformed almost every corner of the biological sciences. Historically, in 1977 Sanger and Coulson developed Deoxyribonucleic Acid (DNA) sequencing technology which was based on chain-termination method (Sanger et al., 1977) and Maxam and Gilbert developed the DNA sequencing method based on the nucleobase-specific partial chemical modification of DNA and subsequent cleavage at specific bases (Maxam and Gilbert, 1977). Sanger sequencing was adopted as the primary technology in the first-generation of laboratory and commercial sequencing applications due to its high-efficiency and low radioactivity (Liu et al., 2012). The second-generation sequencing technologies produce reads length ranging from 35 to 400 bp, at far greater speed and lower cost than Sanger sequencing (Schatz et al., 2010). The second-generation often indicates a platform that requires amplification of the template molecules prior to sequencing while third-generation often indicates platforms that sequence directly individual DNA molecules. NGS platforms generically indicate second or third-generation instruments. Current NGS platforms produce shorter reads than Sanger sequencing (NGS reads are 50 to 150 bp), but with vastly greater numbers of reads, as many as 6 billion per run. Initially using Sanger sequencing platform original human genome project generated approximately 30 million reads, with lengths up to 800 bp and automated capillary sequencers (Lander et al., 2001; Venter et al., 2001). By contrast, the NGS machines can now sequence the entire human genome in a few days with more efficiency. Indeed, DNA sequencing efficiency has increased by approximately 100,000 fold in the decade since the sequencing of the human genome was completed (Treangen and Salzberg, 2012). This also motivated the entomologists for launching an ambitious initiative to sequence 5,000 insect genomes in a project known as i5k. The currently available high-throughput sequencing (HTS) technologies include Illumina

(Hiseq2500, Miseq), Ion Torrent (PGM 318, Proton I), Applied Biosystems (ABI) SOLiD, Pacific Biosciences Real-time Sequencer (PacBio RS), Roche 454 (FS FLX+, GS Junior) and Helicos Heliscope Genetic Analysis System.

Nevertheless, generating large, continuous regions of DNA sequence by the high-quality assembly of reads is still challenging mainly because of repetitive sequences. In addition to creating gaps in assembly, repeats can be erroneously collapsed on top of one another and subsequently can cause complex, misassembled rearrangements (Phillippy et al., 2008; Pop and Salzberg, 2008). Moreover, during the assembly sequence reads are often incorrectly discarded as mistakes or repeats, and others are joined up in the wrong places or orientations due to the abundant repetitive sequences (Baker, 2012). In the presence of reference genome, the comparative assembly is used where the reads are aligned to the reference genome in order to characterise the test genome. However, in the absence of reference genome de novo assembly is used to reconstruct the novel genome that mathematically falls within NP (nondeterministic-polynomial time)-hard problem for which no efficient solution is known (Myers, 1995). Notably, both the approaches are not exclusive as even in the presence of reference genome, unaligned and significantly different regions of the query genome can only be reconstructed through de novo assembly (Pop, 2009). In comparative and functional genomics, the misassembled genome further may lead to wrong interpretation of absence or presence of the genes. Herein the conjunction with HTS transcriptome data helps in re-assembly and in re-annotation process.

1.2 Transcriptome sequencing

Transcriptome study with RNA-Seq is one of the appealing application supported by NGS platforms Depending upon the experimental goals, the NGS platform can be carefully chosen for transcriptome sequencing. The five sequencing platforms (Illumina HiSeq, Life Technologies PGM and Proton, Pacific Biosciences RS and Roche 454) showed a high correlation coefficient (Spearman rank $R > 0.83$) for normalized transcript expression measures across the deep-count platforms (Li et al., 2014). In general, all the sequencing platforms are often producing high-quality, consistent data despite different underlying technologies, both for library preparation and the sequencing itself provide the evidence that RNA-Seq is reliable and reproducible technology. Illumina and SOLiD are typically characterised by low sequencing error rate in compare to single-molecule based sequencing platforms such as Helicos. Lowly expressed transcripts can also be detected by depth-higher sequencing capacity of Illumina and SOLiD. Roche 454 and PacBio platforms are good for

sequencing of longer reads, although the paired-end (data describe both 3′ and 5′ ends of the original Ribonucleic Acid (RNA) species prior to amplification) sequencing approach implemented by Illumina enables it to provide sequence information for a read that is a few hundred nucleotides long (Chu and Corey, 2012).

RNA-Seq experiments can serve to address many biological issues such as quantification of genes, transcriptome profiling, the pattern of alternative splicing and allele-specific expression of transcripts with an unprecedented level of sensitivity (Costa et al., 2010). In immunological studies, RNA-Seq is particularly advantageous in assessing the current state of a cell or tissue and the effect of immune challenge on the transcriptome. Although microarrays experiments fueled the transcriptomics and proved effective in determining gene expression profiles, RNA-Seq, by comparison, is more sensitive, provides absolute quantity levels, is not affected by on-chip sequence biases, and gives additional information on gene expression levels and splice junction variants (Bryant et al., 2012; Labaj et al., 2011). Moreover, one of its important applications of RNA-Seq data the improvement of existing genome annotations (i.e., re-annotation) in eukaryotes.

1.3 Genome re-annotation

In general, the re-annotation can be defined as the improved structural and functional annotation of previously annotated genome by integrating the biological expertise, advanced computational analysis and auxiliary biological data, for instance, express sequence tags (ESTs) and reads generated by transcriptome sequencing. The genome annotation could be erroneous because of several facts such as an error in gene models due to the wrong prediction of splicing junctions, erroneous and inconsistent gene naming due to the transferred annotation based on Blast (Altschul et al., 1990) similarity where the original gene name is itself incorrect (Devos and Valencia, 2001). Additionally, the flawed functional annotation is inevitably propagated by Blast based annotations if not followed by the extensive and careful manual curations. The transcriptome based re-annotation have also been performed on extensively used model organisms such as *Drosophila melanogaster* (Misra et al., 2002) and *Rattus norvegicus* (Li et al., 2015). The re-assembly and re-annotation procedure has shown its importance in improving the improper original annotations, for instance, the recent update of *A. mellifera* genome resulted in the annotation of approximately 5000 more protein-coding genes than previously reported (Elsik et al., 2014). In the comparison of expressed sequence tag (EST), RNA-Seq provides comprehensive data for the discovery of novel genes and transcripts, which is instrumental in gene discovery and gene sequence determination (Li

et al., 2015). Along with the EST/cDNA, assembled or unassembled RNA-Seq reads can be mapped to reference genome using aligners such as Tophat2/Bowtie2 (Kim et al., 2013; Langmead and Salzberg, 2012) to generate the evidence of splice sites, introns, and exons. These evidences can be integrated into gene predictors for improving existing gene structure annotations and novel gene discovery. More coverage of genome by the transcriptome data eventually increases the reliability of the annotation. The mapping of transcriptome sequence reads on the corresponding genome also provides the clue of exon-intron boundaries which could also be exploited for correct annotation of isoforms. The robust gene predictors are quite flexible to incorporate data also from other sources including ESTs and correctly annotated proteins. In this study, the transcriptome data is utilized for re-annotation of an insect *C. floridanus* genome, identification of alternative transcripts and detecting the differentially expressed immune genes.

1.4 The insect immune system

The adaptive immune system evolved in the ancestor of the vertebrate lineage while the innate immune system evolved early in the evolution of multicellular life (Cooper and Alder, 2006). Despite the lack of adaptive immunity, insects have been widely used to study the immune system complexities partially due to the similarity of the innate immunity components with human (Vilmos and Kurucz, 1998). Insects have a multilayered defence system that protects them from different pathogens. The insect defence system also comprises the physical barriers, together with local and systemic responses against broad classes of pathogens (Uvell and Engstrom, 2007). Depending on the route of the transmission of pathogen the different physical barriers ensures the protection (Vallet-Gely et al., 2008).

The production of reactive oxygen species (ROS) through the NADPH dual oxidase (DUOX) by epithelial innate immune system constitutes the strong inducible insect defence. ROS-dependent protection against pathogen involves in the direct killing of invading pathogens as one of the primary layers of resistance (Kim and Lee, 2014). Moreover, the lack of adaptive immune system in insects is compromised by the presence of robust and innate immunity that allows a general and rapid response to infectious agents (Beckage, 2011; Cooper and Alder, 2006). The innate immunity refers to the intrinsic ability of an organism to detect and counteract the potentially harmful activities of nonself foreign agents (Janeway and Medzhitov, 2002). The immune system of insects is endowed with humoral and cellular responses as a part of an innate immune system that work together to combat with infections (Marmaras and Lampropoulou, 2009). Notably, considerable overlap and crosstalk occur

between humoral and cellular defences and the effectors involved in one process could be used by other processes, for instance, the thioester-containing proteins (TEPs) are often produced by the Jak-Stat humoral immunity pathway but they participate in phagocytosis which is considered as a process that underlies in cellular immune defences.

The humoral immune system comprises the signalling pathways such as Toll, IMD, Jak-Stat, and JNK pathways, which ultimately lead to the activation of immune responses upon activation by invading microbes (Ferrandon et al., 2007; Leulier and Lemaitre, 2008). The pattern recognition receptors (PRRs) of these pathways recognise the microbe-associated molecular patterns (MAMPs) of the microbial intruders and initiate the signalling cascades. Subsequently, the mounted humoral immune response leads to the production of antimicrobial peptides (AMPs) (Bulet et al., 2004; Haine et al., 2008) to combat the microbes and numerous other genes as cytoskeletal genes, stress-induced genes, and other effectors. Invading microorganisms that break the primary passive protective barriers such as the cuticle and peritrophic membrane in the gut are encountered by immediate-acting defence mechanisms as phagocytosis, phenoloxidase activity, reactive oxygen species (ROS) mediated immunity and other responses. Furthermore, the cellular defences include phagocytosis, nodulation and encapsulation (Schmidt et al., 2001; Strand and Pech, 1995) often occurs in haemolymph and empower the insect defence.

Ingestion of bacteria induces the transcription of AMP coding genes via the signalling pathways, both locally in the gut epithelium and systemically in the fat bodies. Illumina sequencing was used to identify the differentially expressed genes in *C. floridanus* challenged with mix population of Gram-positive and Gram-negative bacteria. Apart from AMP-mediated immunity, several studies in insects have also provided the evidence for the involvement of haemolymph proteins in eliciting immune responses. Hemocoel of insects is filled with the circulating fluid haemolymph consist of water, inorganic salts, free amino acids, multiple proteins and other organic components. For the identification of induced haemolymph proteins during bacterial infection tandem mass spectrometry was performed.

1.5 The genome of ant *C. floridanus*

Ants develop in four distinct life cycle phases: egg, larva, pupa, and adult. Briefly, the life begins with an egg that hatches into a worm-shaped larva which relies on adults for food requirements. After a rapid growth phase when the larva is large enough, it metamorphoses into a pupa. Pupae do not eat food but slowly develop and finally emerges as an adult. In

terms of immunity, studies have shown the differences in the immune responses of larvae and adults of various insects (Colgan et al., 2011; Fellous and Lazzaro, 2011; Randolt et al., 2008).

Together with the *Harpegnathos saltator*, the Florida Carpenter ant *C. floridanus* genome was sequenced in the year 2010 which eventually drove the insect research also towards the ant genomics (Bonasio et al., 2010). As per efforts of insect research communities and advancements in the field of sequencing technologies, fourteen ant genomes have been sequenced within last six years. Ants are eusocial insects belong to the family Formicidae and shows the typical division of labour in the colonies. Besides the sequenced honey bee *Apis mellifera*, ants are also exploited as a model species for studying social behaviour. Nevertheless, the study on ants may also provide insights into longevity, development, metabolism, behaviour, epigenetics, gene regulation and immunity. This study is considerably focused around the immune system of ant *C. floridanus*.

The first report of *C. floridanus* genome sequencing reveals that it has about 240 million base pairs and 17,064 transcripts codes for 17,064 proteins cflo_OGSv3.3 (Cflo.v.3.3). A total of 231 immune genes was reported including 76 humoral immunity pathway genes (Bonasio et al., 2010). In light of the current interest in ant research, improved annotation is essential also to present the better classification on ant immune system. Thanks to HTS platforms to provide the sequenced ant genomes that opened the new horizon for ant research. The genome size of *C. floridanus* is 303-323 Mb as determined by quantitative PCR (qPCR) and the assembly size is 240 Mb (Bonasio et al., 2010).

1.6 Immunity and symbiosis

Recent investigations have shown the evolution of beneficial interactions between insects and microbes. Similar to vertebrates, insects resident gut microflora, albeit of a much lower complexity (Engel and Moran, 2013). Furthermore, often very intimate, long-term and obligate interactions with endosymbiotic mutualistic bacteria have evolved, which are even transmitted vertically to the offspring (Moran et al., 2008; Moya et al., 2008). Thus, like vertebrates, insects require strategies to control invading harmful bacteria, but they must also be able to tolerate or support the useful microflora, if present.

Immune systems play a fundamental role in insects to distinguish between pathogenic and symbiotic bacteria and regulate their immune response accordingly. Interestingly, the differential treatment of pathogenic and endosymbiotic bacteria by the immune system is one of the most intriguing characteristic of insects (Ratzka et al., 2013; Ratzka et al., 2011). The

versatility of the insect to interact with pathogenic and symbiotic bacteria is attributed to their well-orchestrated immune system. The understanding of complex interactions between host-pathogens or host-symbionts at system level could provide valuable insights of host communications with microbes. To study such behaviour globally it is essential to characterise the immune system of insect species which is under investigation. Moreover, how the integrity of the system affects by the modulation of a connector can provide the valuable information. Consequently, network-based approaches are increasingly used to understand the impact of each component on others and are becoming the starting point of many scientific discoveries.

1.7 Protein-protein interactions (PPIs) network

Often many cellular processes are carried out by molecular machines whose action is coordinated through complex networks of protein interactions. Therefore, the inter-relationships between proteins, rather than the individual protein, eventually determine the behaviour of multiple coordinated processes in a biological system. Unfortunately, most large-scale PPIs are only available for of model organisms and the limited number of some non-model organisms, thus because of the role of PPI in the biology of organism systematic inference of interactome is crucial. Although due to advances in high-throughput methods, the amount of PPI data has increased but these methods are highly susceptible to noise. It is thus the computational techniques are used not only to assess the reliability of a PPI but also the transfer of interactions to organisms of interest (Shoemaker and Panchenko, 2007). To complement the experimental techniques, a number of methods have been developed to predict PPIs (Shoemaker and Panchenko, 2007). Interolog method is established as one of the robust methods for PPIs inference, which combines known PPIs from one or more source species and orthology relationships between the source and target species to identify the putative PPIs in the target species (Walhout et al., 2000). Additionally, interacting domains, co-localization, and functional relationships between the pair of proteins is accessed to identify the plausible PPIs (Roslan et al., 2010). The strict co-localization filtering can omit the signalling interactions occurring between proteins localized in different cellular components. Therefore, to predict the signalling interactions the data for interacting signalling proteins can be used to infer the signalogs i.e., the interolog of protein pairs with directionality inferred from signalling interactions in template species (Roslan et al., 2010).

1.8 Host-pathogen and host-symbiont PPIs

The differential expression of immune genes of *C. floridanus* in response to pathogenic bacteria has been investigated by suppression subtractive hybridization (Ratzka et al., 2011) and Illumina HiSeq sequencing of the immune challenged transcriptome (Gupta et al., 2015). However, these studies do not explore the contributor proteins from pathogen side involved in interactions with an ant immune system. Indeed, the protein-protein interactions (PPIs) predictions can contribute here to relatively under-explored field of insect-bacteria interactions and opens new avenues for insect biologists. There are few shreds of evidence that the carpenter ant immune system has a role in generating host-pathogen specific response. Previous work in our laboratory reported increased differential gene expression of ant immune genes and AMP hymenoptaecin in response to Gram-negative and Gram-positive bacteria (Ratzka et al., 2012). Although in vivo experiments with insects have revealed the essential role of insect immune genes required for maintenance of the endosymbiont number in the bacteriome and ensure endosymbiont persistence within the host tissues (Anselme et al., 2008; Kremer et al., 2012; Ratzka et al., 2013; Stoll et al., 2009), the global scenario of insects and endosymbiont interactions is still needs to be investigated.

1.9 Mass spectrometry analysis of haemolymph

Synthesis of RNA and specific proteins in the haemolymph render the increase in antibacterial immune responses (Boman and Hultmark, 1987). Notably, some factors relevant to immunity are normally present in the haemolymph, for instance, prophenoloxidase and lectins. However, most proteins in haemolymph are secreted from other tissues therefore, quantification of protein during the immune challenge is more restricted to proteomics. In this scenario, Mass spectrometry (MS) is a robust method for large-scale protein identification from complex mixtures of biological origin (Aebersold and Mann, 2003; Baldwin, 2004). In MS-based proteomics, Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF-MS) method is often used to get MS/MS spectra intact proteins based upon the comparison of the protein fingerprint obtained from the test sample with a theoretical MS/MS spectra generated from an in-silico digested database of the known proteins of specified species. Generally, combining proteomic datasets with complementary transcriptome data can reveal interesting facets of cellular functions, for instance, the key regulators of biological process (Kumar and Mann, 2009). However, the lack of correlation between mRNA and protein levels is well-documented and has been attributed to differences in translational efficiency, codon usage/bias, and mRNA versus protein stability (Ghazalpour

et al., 2011; Gygi et al., 1999). One of the major contributors that affect the correlation between mRNA and protein levels is miRNA which can affect the gene expression via either translational repression or mRNA degradation at the post transcriptional level. Although the mRNA expression data used here comes from Illumina sequencing of whole animal and MS was performed specifically to haemolymph, the work presented here tried to correlate both the data and also identified if some of the genes differentially expressed during immune challenge are the direct targets of miRNA.

2

Materials and Methods

This chapter summarizes the research methods of this thesis. It is worth to mention here that all the wet laboratory experiments were performed by my co-workers in the Microbiology Department, but for the sake of the completeness of the thesis also presented here. Part I defines the experiments (performed by Carolin Ratzka and Maria Kupper, Department of Microbiology, University of Würzburg) and part II describes the bioinformatics methods. The data from transcriptome sequencing were used for re-annotation, analysis of differentially expressed genes and mapping on interaction networks to identify functional modules, as elucidated in bioinformatics methods. The data from haemolymph proteome quantifications were used for enrichment analysis, mapping on interaction networks and observation of stage specific differences by bioinformatics methods.

Part I experimental

2.1 RNA sample preparation and extraction for Illumina sequencing

Late larvae (stage L2) and adult minor workers (stage W2) of *C. floridanus* colonies were pricked with a minutiae needle (Minutiennadel Sphinx V2A 0.1 x 12 mm, Bioform), which was previously dipped into a pellet of mixed heat-killed bacteria (1:1proportion of *Escherichia coli* D31 and *Micrococcus luteus*). The stage definitions are termed elsewhere (Stoll et al., 2010). After 12 h of immune-challenge total RNA of 5 pricked and 5 non-inoculated ants were extracted using TRIzol® Reagent (Invitrogen, Carlsbad, CA, USA) and purified through RNeasy Mini Kit columns (Qiagen, Hilden, Germany) with on-column DNase digestion (RNase-Free DNase Set, Qiagen). The concentration, integrity and quality of total RNA were determined on an Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Chip kit (Agilent Technologies, Böblingen, Germany). Afterward, equal amounts of total RNA from immune-challenged workers and larvae as well as each 5 µg from untreated L2 and W2 were mixed and the two resulting RNA samples were further processed

and sequenced with Illumina HiSeq2000 (2 x 50 bp paired-end sequencing) by Eurofins MWG Operon (Ebersberg, Germany).

2.2 Validation of differential gene expression results by qRT-PCR

qRT-PCR experiments were performed for the validation of differential gene expression results obtained by transcriptome analysis. The details for the sample and experimental setup is described elsewhere (Gupta et al., 2015). Briefly, RNA samples were isolated from treated as well as from untreated animals and the differences in the expression of the 37 differentially regulated genes, 12 hours after the post infection was analysed by qRT-PCR separately regarding L2 and W2. The qRT-PCR experiments were performed on a StepOnePlus™ Real-Time PCR System (Applied Biosystems, Life Technologies GmbH, Darmstadt, Germany). Statistica v10 enterprise x64 was used for statistical analysis. The t-test was used to test whether the relative gene expression of six biological replicates of infected animals (dCT Immune) differs significantly from, relative gene expression of control animals (dCT Control).

2.3 Immune-challenge, haemolymph isolation and sample preparation for mass spectrometry

For immune-challenge a minutiae needle (Minutiennadel Sphinx V2A 0,1x12mm, Bioform, Nuremberg, Germany) was dipped into a bacterial pellet of mixed heat-killed bacteria (1:1 proportion of *E. coli D31* and *M. luteus*) and adult worker ants were pricked between the first and the second segment of the gaster while larvae were pricked ventrally on the level of the midgut. For haemolymph sample collection of control and infected larvae were picked with a minutiae needle or a pulled glass capillary on the lateral side directly behind the head and a micro capillary pipette was held close to the wound while the animal was slightly pressed with forceps. Moreover, in case of adult worker ants head, thorax and abdomen were separated using dissecting scissors and by applying slight pressure on head and thorax haemolymph was squeezed out of the body parts and collected with a micro capillary pipette. To avoid the haemolymph melanisation, the extracted haemolymph samples were held on ice and a few microliters of a mixture of aprotinin and N-phenylthiourea (each 0.2 mg/ml, both from Sigma-Aldrich) were added. Samples were centrifuged at 14,000 x g for 1 min at 4°C to remove haemocytes and cellular debris. For precipitation of peptides from proteins within the haemolymph samples, acetone was applied to each sample in a 1:5 ratio and samples were incubated at -80 °C for 2 hours. Furthermore, the separation was achieved

by centrifugation at 14,000 rpm at RT. The protein samples (pellets) were provided to the cooperation partners and further prepared for MALDI-TOF-MS analysis.

Part II bioinformatics

2.4 Assembly, detection of new transcripts, and differential gene expression analysis

The 2 x 50 bp paired-end sequence reads from both the sample were mapped onto the *C. floridanus* genome using TopHat (Trapnell et al., 2009) and further the expressed transcripts were assembled using Cufflinks (v2.0.2) (Trapnell et al., 2012). The computations were performed on HP ProLiant DL580 G5 offering four Intel(R) Xeon(R) CPUs E7440 and 40 GB of RAM (performed by Frank Förster, Department of Bioinformatics, University of Würzburg). The merged transcriptome annotation was generated using Cuffmerge and differentially expressed genes (DEGs) were identified by two independent methods namely Cuffdiff (Trapnell et al., 2012) and DESeq (Anders and Huber, 2010). Cuffdiff calculates transcript differential expression based on their respective transcript abundances in control and treated samples while DESeq works on count data for differential expression analysis and attempts to identify if the observed difference between two conditions can be attributed significantly due to the experimental condition alone and not due to the biological variance.

2.5 Identifying repetitive elements

The *de novo* repeat library customized for *C. floridanus* genome was constructed with RepeatModeler (Smit and Hubley, 2011). The repeats for *C. floridanus* genome present in Repbase library (Jurka et al., 2005) were integrated with *de novo* repeat library. The combined library was further used to delineate various interspersed repeats and low-complexity regions present in *C. floridanus* genome and subsequently masked with RepeatMasker.v4.0 (Smit et al., 1996-2013).

2.6 Structural and functional re-annotation

The extended version of Generalized Hidden Markov Model (GHMM) based *ab initio* predictor Augustus.v2.7 was used for reconstructing gene models. The high-quality training set consisting of 330 genes was constructed for the generation of species-specific parameters for the splice site signals, nucleotide composition of exons, length distributions, introns and intergenic regions. Cegma.v2.4 (Parra et al., 2007), PASA2 release 2013-06-05

(Haas et al., 2003) and Scipio.v1.4 (Keller et al., 2008) were used to derive the high-confidence gene models for the training set. The HMM training and retraining was performed to calculate and optimize the gene predictions parameters. The flexibility of Augustus allows integrating the extrinsic evidence (hints) for more accurate gene predictions. The sources of hints include (a) the raw Illumina sequencing reads (b) the assembled transcripts by Cufflinks and (c) all available EST data of *C. floridanus*. Augustus predictions with hints were performed on the repeat-masked genome using the tuned optimized parameters of *C. floridanus* to predict the gene structures and estimate alternative transcripts. The functional annotation of predicted proteins from re-annotation was performed with AutoFACT.v3.4 (Koski et al., 2005), Blast2GO (Conesa et al., 2005), and Pfam database (Finn et al., 2013).

2.7 Identification of immune genes

Orthology-based annotation, Gene Ontology (GO) analysis, exhaustive analysis of literature and databases were performed to identify the immune related genes and the corresponding proteins in *C. floridanus*. Immune genes of *D. melanogaster*, *A. mellifera*, and *Nasonia vitripenni*s were collected from the previously published data and their orthologs were identified in *C. floridanus* using OrthoMCL.v2.0.9 (Li et al., 2003). GO analysis was performed with Blast2GO. Moreover, to detect the presence of potential immune effectors including chitinase, lysozymes, prophenoloxidase, nitric oxide synthase, glutathione S-transferase, TEPs and turandots, BlastP search was implemented using the effector sequences from several insects (*D. melanogaster*, *A. mellifera*, *Anopheles gambiae*, *Bombyx mori*, *Manduca sexta*, *Aedes aegypti* and *Phaedon cochleariae*) as query. All the genes were manually curated and categorized into multiple groups based on their immune functions.

2.8 Reconstruction of immune signalling pathways

The list of immune genes involves in insect humoral immunity pathways (Toll, Jak-Stat, IMD, and JNK) were collected from research articles, textbooks, and electronic information sources. The directed interactions between the pathway components were retrieved from three pathway databases KEGG (Kanehisa and Goto, 2000), FlyReactome (Matthews et al., 2009) and INOH (Yamamoto et al., 2011) followed by manual curations from published literature. The proteins and connectivity information were translated into comprehensive immune signalling networks of *D. melanogaster*. Homologs of these proteins in *C. floridanus* were determined by BlastP (Altschul et al., 1997) based similarity search against *C. floridanus* proteome. Pfam database was used to analyse the protein domains in

homologs and the function of homologs was assigned by domain conservation between the homologs. The identified homologous immune proteins were mapped onto reconstructed immune signalling networks of *D. melanogaster* and pathway annotations were transferred if any of the two immune proteins in *C. floridanus* had corresponding interacting homologs. Employing the same strategy, the reconstructed *C. floridanus* signalling networks were further extended by the mapping the additional members of humoral immunity pathways present in *A. mellifera* but absent in *D. melanogaster*.

2.9 Identification of antimicrobial peptide (AMPs)

HMMsearch module of HMMER3 (Mistry et al., 2013) with an e-value threshold of 1e-03 was used to scan all the protein isoforms obtained from re-annotation of *C. floridanus* against HMMs of AMPs retrieved from AMPer database (Fjell et al., 2007). Additionally, the AMPs of other insect were used as the reference set to identify additional AMPs in *C. floridanus* based on BlastP and TblastN searches followed by manual curations of Blast hits.

2.10 Orthology analysis

Seven other sequenced ant genomes, including *A. cephalotes* (leafcutter ant), *A. echinatior* (Panamanian leafcutter ant), *P. barbatus* (red harvester ant), *H. saltator* (Jerdon's jumping ant), *L. humile* (Argentine ant), *S. invicta* (red fire ant), *C. biroi* (clonal raider ant) and three other insects, including a model insect *D. melanogaster* (fruit fly), model social insect (honey bee), and a solitary insect *N. vitripennis* (parasitic wasp) were chosen to examine the conservancy of *C. floridanus* immune proteins. OrthoMCL was used to establish the orthology relationships.

2.11 Inferring PPIs and signalome of *C. floridanus*

The proteome-scale experimentally verified high-confidence PPIs in *D. melanogaster* was retrieved from the multiple sources including interactions from LexA Y2H system screens (Schwartz et al., 2009; Stanyon et al., 2004; Zhong et al., 2003), high throughput Gal4 proteome-wide yeast two-hybrid (Y2H) screens (Giot et al., 2003), PPIs from *D. melanogaster* protein interaction map (Formstecher et al., 2005) and interactions determined in large-scale co-affinity purification (co-AP)/MS screens (Friedman et al., 2011; Guruharsha et al., 2011). Moreover, to avoid the biasedness towards *D. melanogaster* interaction data the PPIs from BIND (Bader et al., 2003), BioGRID (Stark et

al., 2011), MINT (Licata et al., 2012), IntAct (Orchard et al., 2014), and Database of interacting proteins (DIP) (Salwinski et al., 2004) were also used as reference set for mining *C. floridanus* PPIs. To determine the interologs two robust orthology prediction algorithms i.e., OrthoMCL and InParanoid.v.4.1 (Östlund et al., 2010), and customized in-house developed Perl and Bash scripts were used. Blast e-value 1e-05 and Markov Clustering (MCL) inflation index 1.5 was used for OrthoMCL analysis while default parameters of InParanoid were used to determine orthologs of DIP interactors however for the *D. melanogaster* data orthology Blosum80 matrix was selected instead of default Blosum62. Besides the seed orthologs identified by InParanoid, the consensus prediction of InParanoid and OrthoMCL were used to create the set of interologs. Functional domains were assigned to the proteins present in preliminary ant interactome using Pfam version 27.0 (Finn et al., 2013). Domain interactions data collected from three different databases namely Domine (Yellaboina et al., 2011), DIMA 3.0 (Luo et al., 2011) and IDDI database (Kim et al., 2012) and used to filter the ant PPIs not supported by domain-domain interactions (DDIs). Furthermore, filtered interactors were assigned to subcellular locations using a combination of KnowPredsite (Lin et al., 2009) extension UniLoc and SwissProt orthology. If both the proteins in an interaction do not share the same localization, the interaction between them was treated as probable nonfeasible interaction and removed from the ant interactome. Finally, in the 'Interactome of *C. floridanus*' (IC) the alternative spliced form of proteins coded by a single gene were represented by a single node, if the alternative forms were found to make similar interactions.

Moreover, as the IC reconstructed here is highly constrained with localization, the signalling interaction that occurs along two compartments can be neglected. Therefore, the signalling networks of *C. floridanus* were reconstructed using signalog method and literature followed by extensive manual curations. The directed interaction data of *D. melanogaster* from SignaLink database (Fazekas et al., 2013) were collected and orthologs of them in *C. floridanus* were mapped. Signalling maps of humoral immunity pathways were generated from literature and homologs were mapped using BlastP. Moreover, the large RNAi screens based large signalling map of *D. melanogaster* were retrieved from SignedPPI database (Vinayagam et al., 2014) and similarly the orthologous directed interaction in *C. floridanus* were mapped. Similar to IC, the isoforms were represented by a single identifier. The computations for signalog identification were performed by InParanoid, customized in-house

Perl and Bash scripts. The three sources of signalling interactions were merged to reveal the first draft of the 'Signalome of *C. floridanus*' (SC).

Finally, both the interactome and signalome were integrated to reconstruct 'Integrated Interactome and Signalome of *C. floridanus*' (IISC). The mRNA expression data gained by Illumina based transcriptome sequencing was mapped over the IISC to obtain the infection induced network.

2.12 Network analysis and visualization

Topological analysis of IC, SC and IISC was performed using Network Analyzer plugin of Cytoscape version 2.8.1 (Shannon et al., 2003). The number of hubs (highly connected nodes), degree (connectivity) of nodes, the network diameter (the maximum of the shortest path lengths) and the mean path length (the average of the shortest path lengths) were determined with graph theoretical analysis. Cytoscape was used to visualize the ant interactome and subnetworks. Illumina data was mapped on IISC network to identify the differentially expressed hubs and bottlenecks. Hubs were defined as nodes with > 5 signalling or interacting partners. Top 20 % of differentially expressed bottlenecks were considered important for IISC during bacterial infection.

2.13 Infection-induced IISC-subnetwork construction

DEGs identified from Illumina sequencing were set as seed nodes. These nodes were marked in IISC network and shortest paths connecting them were identified using Dijkstra's algorithm (Skiena, 1990). Afterwards, all the paths found for all pairs of seed node were merged to derive the active IISC-subnetwork during bacterial infection. The nodes in the network were grouped by their functional annotation. Moreover, differentially expressed hubs and bottlenecks were also mapped in IISC-subnetwork to identify the major regulators of induced subnetwork during the immune challenge.

2.14 *C. floridanus-S. marcescens* and *C. floridanus-B. floridanus* PPIs

Experimentally verified host-bacteria interactions were collected from three databases PHISTO (Tekir et al., 2013), PATRIC (Wattam et al., 2013) and HPIDB (Kumar and Nanduri, 2010) and retrieved a total of 8677 non-redundant PPIs shared between 18 eukaryotic hosts and 96 bacterial pathogens. This combination of template data of interactions is further referred as eukaryotic-bacteria interaction (EBI) set. These interactions were used as a template to determine host homologs of *C. floridanus*, and bacteria homologs of

pathogen *S. marcescens* and endosymbiont *B. floridanus* using InParanoid and OrthoMCL. The functional domains in the template set was annotated with Pfam and feasible interactions between domain were identified by DDIs reference set to classify the EBI set into DDI supported and DDI not supported interactions using customized Perl scripts. The predicted ant-bacteria was constrained in such a way that the interolog of DDI supported template should also have DDI support while the interolog of DDI not supported template do not need to have DDI support. In the case of host-pathogen interaction the pathogen interference with the host, components were focused while in the case of the ant-endosymbiont network the evolutionarily conserved interactions were focused. The conservancy of *C. floridanus - B. floridanus* interaction pairs were evaluated in five other pairs of insect symbionts which include *Acyrthosiphon pisum - Buchnera aphidicola, Pachypsylla venusta - Carsonella ruddii, Rhodnius prolixus - Rhodococcus rhodnii, Glossina brevipalpis - Wigglesworthia glossinidia and D. melanogaster - Wolbachia pipientis wMel* species pairs and based on conservancy of interaction, score was given to interaction pairs of *C. floridanus - B. floridanus*. Furthermore, the interaction of *S. marcescens* with IC was analysed to understand how the pathogenic bacteria interfere with the ant immune system.

2.15 Biological filters

The interactors from the bacterial side were selected based on additional evidence for plausible interactions such as membrane localization, orthology with virulence-associated proteins, lipoprotein, secretory and non-classical secretory signal containing proteins. The immune genes of host carpenter ant were systematically collected from own research article (Gupta et al., 2015), extracting functional information from orthologs of genes in different insects and BlastP search against multiple customized Blast databases of insect proteins having role in insect defence system such as gut epithelium proteins, mucins, opsonins and others (Vallet-Gely et al., 2008). The ant proteins in the interspecies interactome were further manually curated to mark the components that may have an important role in insect immunity. Both the host-bacteria interaction networks were created and analysed with Cytoscape.

2.16 Gene Ontology (GO) annotations

Functional enrichment analysis was performed with Blast2GO suite to annotate the biological process, molecular function and cellular components of the *C. floridanus* proteins present in the previous and re-annotated version of ant proteome. Fisher's Exact Test (FET)

was applied to evaluate biases in the differentially expressed and bacterial interacting ant protein datasets.

2.17 MS data analysis and GO annotations

Using the standard workflow (Cottrell, 2011) haemolymph proteins were identified and characterised using tandem mass spectrometry (MS/MS) data. The data from the experiments on larvae and adults were processed and analysed together with MaxQuant software (Cox and Mann, 2008) by the cooperation partner (Jens T. Vanselow, Rudolf Virchow Center for Experimental Biomedicine, University of Würzburg). The haemolymph MS/MS data were pre-processed and the identifiers of the proteins made consistent, by mapping the *C. floridanus* re-annotation identifiers over the old *C. floridanus* accession numbers. GO annotation was performed with Blast2GO suite for annotating the biological process, molecular function and cellular compartment of the genes present in haemolymph data. Moreover, the Fisher's Exact Test was applied to identify the statistically significant enriched GO terms. The proteins detected from the haemolymph of larvae and workers were compared. Significantly up-regulated and down-regulated genes detected from MS/MS data were also enriched for the identification of immune-related proteins. Immune related proteins were annotated based on the classification presented in this thesis, additional literatures, and localization of proteins.

Availability of supporting data

Additional data for the re-annotation part can be downloaded from the web link www.bioinfo.biozentrum.uni-wuerzburg.de/computing/Camponotus. These files are: *C. floridanus* Augustus gene annotations (.gff format), *C. floridanus* transcriptome gene coordinates (.gff format), *C. floridanus* transcripts for the translated regions (.fasta format) and *C. floridanus* predicted proteins (.fasta format). Furthermore, the raw data of the transcriptome sequences (just the Illumina reads, no annotation) are deposited in the NCBI bioproject ID263478 Accession: PRJNA263478. These data have the NCBI accession numbers [GenBank:SRR1609918] for the immune challenged and [GenBank:SRR1609919] for the unchallenged animals.

3

Results

This chapter summarizes the results of experiments and bioinformatics analyses. The results of re-annotation, transcriptome profiling, immumomics and comparative analysis described in this chapter (section 3.1 – 3.11) have been published (Gupta et al., 2016; Gupta et al., 2015) while other sections are unpublished. Some passages (specifically section 3.7 – 3.9) have been quoted *verbatim* from the published BMC Genomics article (Gupta et al., 2015) as I am further owner of the copyright of the article (licence type is CC BY 4.0).

3.1 Re-annotation of *C. floridanus* genome

Full transcriptome of *C. floridanus* was sequenced using the Illumina platform and the immune-challenged and unchallenged ants were compared. Paired-end sequencing generated 125,873,897 reads for infected and 118,142,837 reads for untreated animals. Table 1 summarizes the sequence statistics of Illumina sequencing. To generate the hints for the re-annotation procedure raw reads, assembled reads and ESTs were mapped on repeat-masked *C. floridanus* genome. The repeats were masked using *de novo* repeat library constructed with the RepeatModeler and database of repetitive elements Repbase which resulted in detection of 62 and 12 repetitive elements respectively. With the assembled repeat library, 14.57 % of the *C. floridanus* genome was identified to contain repetitive sequences, consisting of 6.34 % of interspersed repeat elements, 1.70 % of simple repeats, 6.54 % low complexity stretches, and 0.01 % of small RNAs and satellites (Table 1).

Mapping of raw reads on masked genome resulted in preliminary intron hints that subsequently used to generate primary gene models with Augustus and identifying exon-exon junction followed by generation of final 111613 intron hints by mapping reads over the exon-exon junction database derived from primary gene models and intron hints. Mapping of 31582 assembled transcripts over unmasked genome produced 80132 intron hints, 146523 exon hints, and 62531 exonpart hints, in contrast, mapping over masked genome generated

71154 intron hints, 126855 exon hints, and 518762 exonpart hints. Similarly, mapping of 61 ESTs over unmasked genome resulted in 70 intron hints, 33 exon hints, and 85 exonpart hints while ESTs mapping over masked genome produced 57 intron hints, 31 exon hints and 54 exonpart hints. All the generated hints were used together as evidence in support of gene models.

Table 1. Quantitative overview on the transcriptome sequencing data.

(A) Illumina sequencing – quantitative overview			
		Immune challenged	Non-immune challenged
Sequencing		Paired end 2 x 50 bp	
Total number of reads		125,873,897	118,142,837
Median insert length (bp)		176 bp	163 bp
Overall mapping rate		87.4 %	88.6 %
Mapping to *Camponotus* genome		99.45 %	99.37 %
Mapping to *Blochmannia* genome		0.55 %	0.63 %
(B) Repeats distribution			
	Number of elements	Length occupied	Percentage of sequence
LTR elements	662	445584 bp	0.19 %
DNA elements	163	44183 bp	0.02 %
Unclassified	41722	14401396 bp	6.13 %
Total interspersed repeats	-	14891163 bp	6.34 %
Small RNA	26	13938 bp	0.01 %
Satellites	50	9914 bp	0.00 %
Simple repeats	67608	3996593 bp	1.70 %
Low complexity	284104	15360412 bp	6.54 %

Augustus software was trained and re-trained to obtain the species-specific parameters for gene prediction in *C. floridanus*. The prediction accuracy of Augustus with the optimized parameters is listed in Table 2. As the Augustus predictor is one of the robust software package used to annotate splicing, exons, introns and genome features, this is an important extension of this annotation tools for the annotation quality in ant genomes.

Table 2. Accuracy of trained gene prediction parameters on *C. floridanus* test set sequences.

	Base level			Exon level		
Program	Sensitivity (Sn)	Specificity (Sp)	(Sn+Sp)/2	Sensitivity (Sn)	Specificity (Sp)	(Sn+Sp)/2
Augustus	0.953	0.906	0.9295	0.82	0.796	0.808

The re-annotation generated 15,631 protein-coding genes which were comparatively lower than 17,064 protein-coding genes reporting previous annotation Cflo.v.3.3 (Bonasio et al., 2010). Despite counting fewer genes, the revised annotation significantly reduces the over-prediction (false positives) of single exon genes and subsequently displayed an increase in the number of multi-exon genes as compared to Cflo.v.3.3. The genes predicted with the optimized species-specific parameters were based on the data from *C. floridanus* ESTs and the large-scale Illumina sequencing data (raw reads and assembled reads). The re-annotation showed the improved quality of the annotation as 14,956 (81.41 %) predicted transcripts could now be supported by extrinsic evidence such as introns, exons, and exon part evidences generated by sources such as Illumina sequence data and ESTs. In terms of proteins, re-annotation counts a total of 18,369 proteins as compared to 17,064 proteins in Cflo.v.3.3. A key improvement in this revised annotation is better identification of alternative splicing events and differential isoforms using the optimized parameters with 80.8 % exon level accuracy.

Table 3. Summary of genes and their alternative splicing products predicted with Augustus run on repeat masked *C. floridanus* genome.

Number of genes	Count of alternative transcripts
1410	2
346	3
106	4
36	5
15	6
10	7
3	8
2	10
# Total - 1928 genes	# Total - 4666 transcripts

The previous annotation Cflo.v.3.3 showed 7,583 alternative splicing events in 2,538 genes while after improving the incorrect gene models 4,666 alternative transcripts were identified as a product of alternative splicing events in 1,928 affected genes (Table 3). Notably, the Cflo.v.3.3 data contain 17,064 transcripts and 17,064 proteins without distinction between alternative isoforms and thus can not be used to maximum advantage. The re-annotation data of *C. floridanus* reported here can be accessed from the project web repository (http://camponotus.bioapps.biozentrum.uni-wuerzburg.de) and distinguish the alternative splicing

products of the genes with the suffix in the accession number as t1, t2, t3 etc. The exact distribution of alternative transcripts over the annotated genes is listed in Table 3.

Out of 15,631 protein-coding genes, 1928 genes were identified to encode alternative transcripts. The optimized gene prediction parameters for *C. floridanus* (http://bioinf.uni-greifswald.de/augustus/submission.php) allow now better structural annotation of ant genomes and *C. floridanus* in particular e.g., for splicing events. Overall, the integration of the transcriptome data guarantees the basis for various types of analyses, for instance, the reduction of isoforms in PPIs network as explained later in this thesis.

3.2 Improvement of PGRP gene structure

In the re-annotation, one of the important findings was the identification of misannotation of peptidoglycan recognition protein (PGRP). Out of four PGRP present in *C. floridanus*, PGRP-LC was wrongly annotated as PGRP-LE in Cflo.v.3.3. The gene structure of misannotated PGRP-LE was improved during the re-annotation and corrected by the addition of the missing N-terminal sequence (Fig. 1).

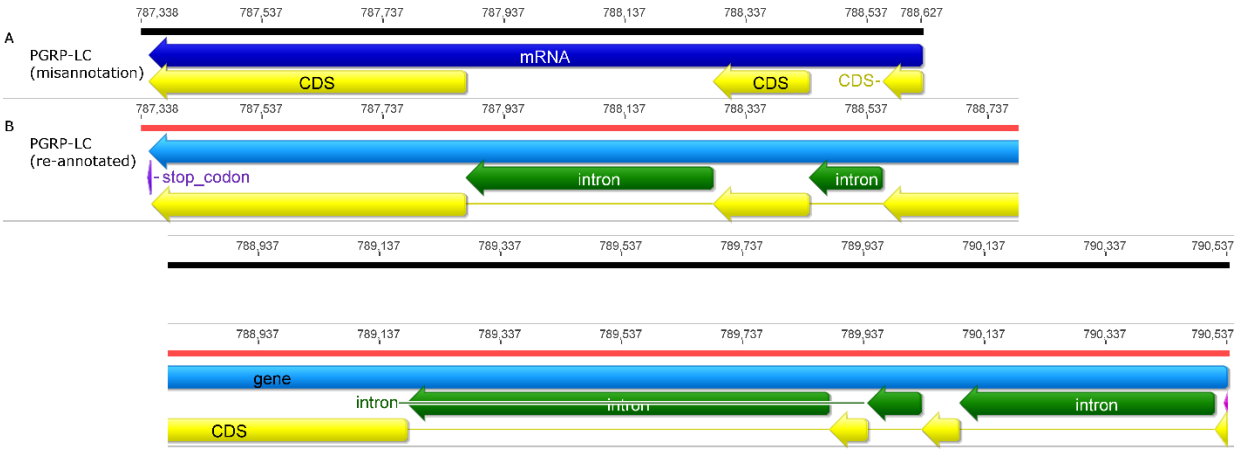


Figure 1: Re-annotation of *C. floridanus* PGRP-LC gene. (A) Original annotation PGRP-LC (wrongly annotated as PGRP-LE; accession number EFN63542.1/Cflo_03358) in Cflo.v.3.3 and (B) Re-annotated version of PGRP-LC (accession number Cflo_N_g10272t1).

The re-annotation parameters determined the presence 6 exons and 5 introns, among which 3 exons and 4 introns were supported by transcriptome based evidence. Including the 3 hints from repeat masked regions, overall 63.6 % of the transcript length of PGRP-LE gene was found to be supported by evidences. The correct name of the protein was adapted from high similarity with *A. mellifera* PGRP-LE during function annotation. It was further confirmed by OrthoMCL clustering of different insects where the older PGRP-LE grouped

within the PGRP-LC cluster. The misannotated PGRP-LE also had a transmembrane domain at sequence position 264 to 287 but originally PGRP-LE protein is an intracytoplasmic protein while PGRP-LC is a membrane-associated protein.

3.3 Functional annotation and GO classification

In comparison to Cflo.v.3.3, the total number of additional proteins identified in re-annotated version was higher but not all of them contained functional domains. The re-annotation represented 978 more functional domains containing proteins that were absent in Cflo.v.3.3 (Fig. 2). The total number of additional proteins identified was higher but not all of them contained functional domains.

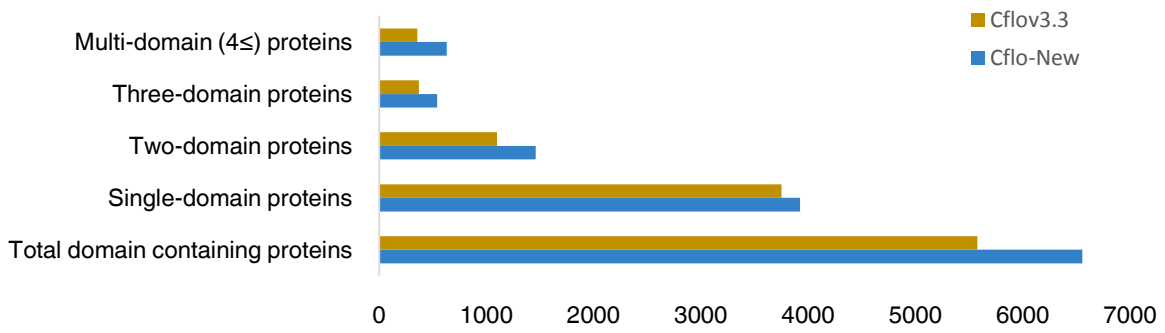


Figure 2: Differences in Pfam protein domains as deduced from the previous and the new *C. floridanus* genome annotation.

The GO annotation of all the proteins in the re-annotated version (Cflo-New) of *C. floridanus* genome resulted in assigned level-2 GO terms to 8490 proteins however, in total 7143 proteins of Cflo.v.3.3 could be annotated by Blast2GO. The comparisons were performed for all the three GO categories, i.e. biological process, molecular function and cellular component.

In the biological process category cellular processes were most abundant (19.98 %) which consist 2709 proteins in re-annotated version while 2248 proteins were enriched with cellular processes in Cflo.v.3.3 (Fig. 3).

In the molecular function, enzymes presented the highest fraction among the annotated proteins (catalytic activity, 40.07 %), followed by the term binding with 39 %. In compare to 2069 enzymes annotated in Cflo.v.3.3, 2440 enzymes were identified in re-annotated version of *C. floridanus* genome (Fig. 4).

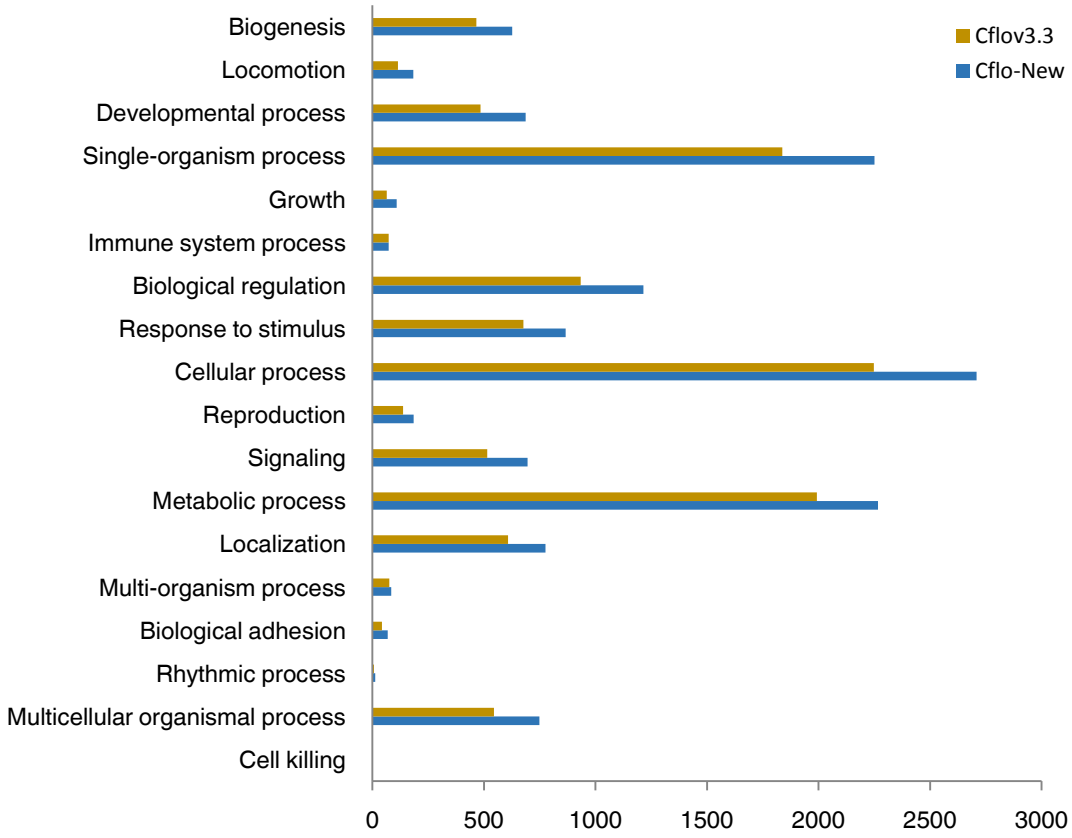


Figure 3: Gene Ontology, biological process category of *C. floridanus*. Previous (Cflo.v.3.3) and new (Cflo-New) level 2 GO functions for *C. floridanus* proteins. The x-axis represents total number of proteins assigned with the given level2 GO term.

Moreover, regarding cellular compartment category term cell was found to be most abundant (33.74 %). However, consider that two thirds of the proteins are located in membranes, organelles or part of macromolecular complexes (Table 4 shows in detail all other functional categories). Comparing with the cellular compartment identified in Cflo.v.3.3, the increase in the membrane protein from 831 to 1024 was also identified in re-annotated version (Fig. 5).

Comparatively, using the same Blast2GO parameters 7143 proteins could be annotated with GO terms in Cflo.v.3.3. In summary, comparing new annotation with Cflo.v.3.3 in terms of GO classification, an increase of 18 to 23 % was reported regarding subcategory assignments covering all three major GO categories (13,354 terms in comparison to 10,819 terms in Cflo.v.3.3 for biological process; 6088 terms versus 5141 in molecular function, and 5441 terms versus 4568 terms respectively for the GO category cellular component (Gupta et al., 2015) (Table 4).

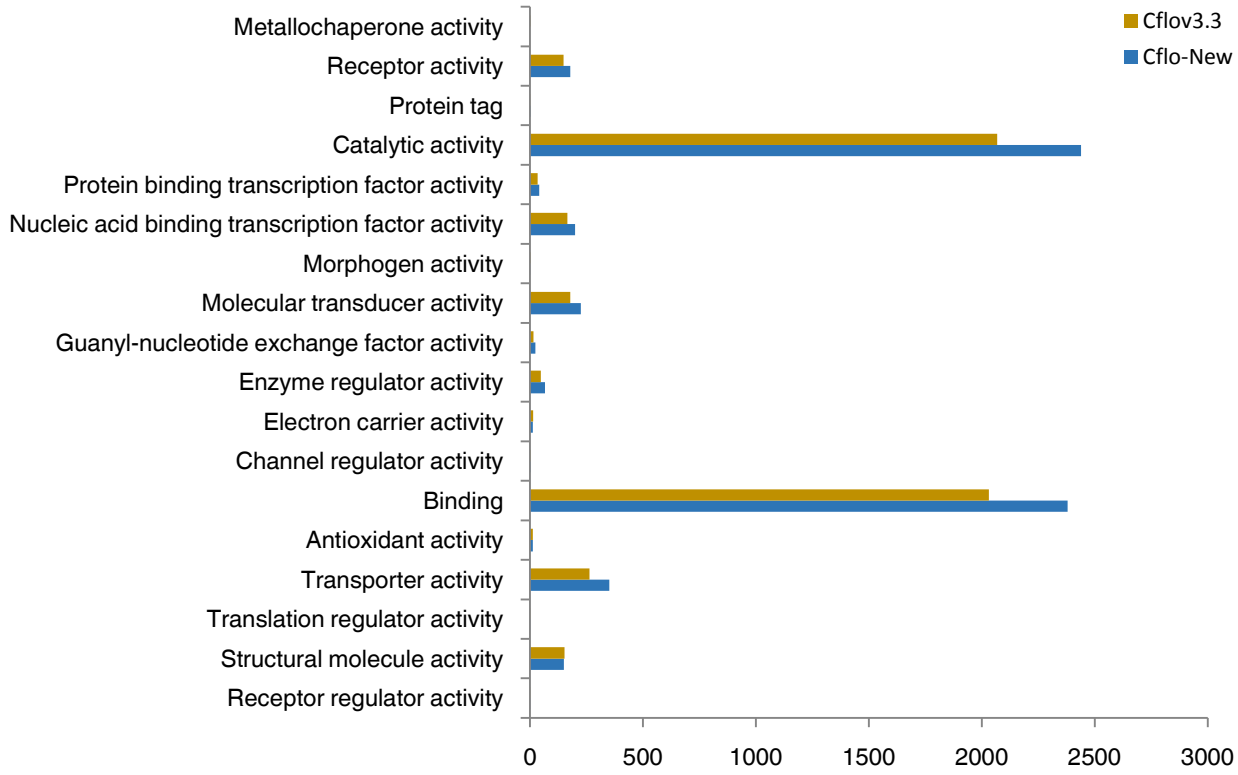


Figure 4: Gene Ontology, molecular function category of *C. floridanus*. Previous (Cflo.v.3.3) and new (Cflo-New) level 2 GO functions for *C. floridanus* proteins. The x-axis represents total number of proteins assigned with the given level2 GO term.

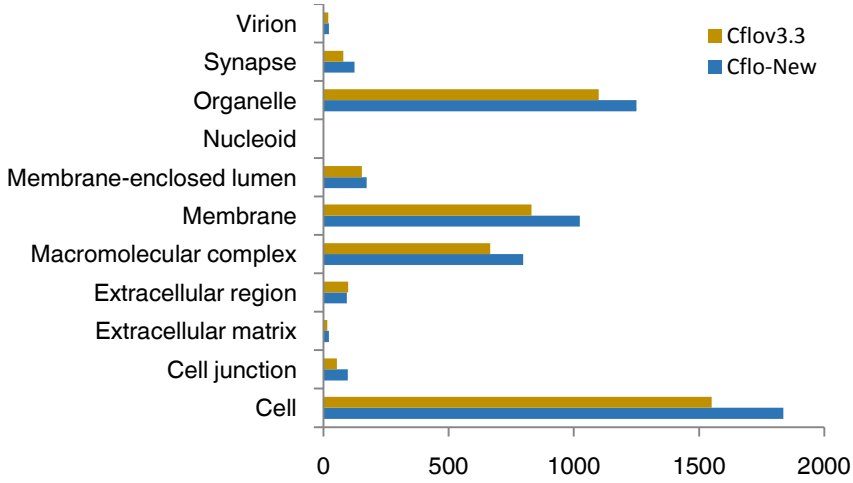


Figure 5: Gene Ontology, cellular compartment category of *C. floridanus*. Previous (Cflo.v.3.3) and new (Cflo-New) level 2 GO functions for *C. floridanus* proteins. The x-axis represents total number of proteins assigned with the given level2 GO term.

Table 4. Global changes in level 2 GO class assignment of re-annotated and Cflo.v.3.3 proteins for GO categories (A) biological process, (B) molecular function and (C) cellular component.

(A) Biological process (Level 2 GO annotation)		
	Cflo-New (13354 terms)	Cflo.v.3.3 (10819 terms)
Cell killing	2	2
Multicellular organismal process	748	545
Rhythmic process	12	7
Biological adhesion	68	42
Multi-organism process	84	75
Localization	775	608
Metabolic process	2267	1992
Signalling	695	515
Reproduction	185	137
Cellular process	2709	2248
Response to stimulus	866	676
Biological regulation	1215	933
Immune system process	73	73
Growth	108	64
Single-organism process	2251	1838
Developmental process	686	484
Locomotion	183	114
Biogenesis	627	466

(B) Molecular function (Level 2 GO annotation)		
	Cflo-New (6088 terms)	Cflo.v.3.3 (5141 terms)
Receptor regulator activity	1	1
Structural molecule activity	150	153
Translation regulator activity	1	1
Transporter activity	351	263
Antioxidant activity	13	12
Binding	2380	2031
Channel regulator activity	3	2
Electron carrier activity	13	14
Enzyme regulator activity	66	48
Guanyl-nucleotide exchange factor activity	24	15
Molecular transducer activity	225	178
Morphogen activity	1	3
Nucleic acid binding transcription factor activity	199	166
Protein binding transcription factor activity	41	34
Catalytic activity	2440	2069
Protein tag	2	1
Receptor activity	178	148
Metallochaperone activity	0	2

(C) Cellular component (Level 2 GO annotation)		
	Cflo-New (5441 terms)	Cflo.v.3.3 (4568 terms)
Cell	1836	1550

Cell junction	98	54
Extracellular matrix	22	16
Extracellular region	93	99
Macromolecular complex	798	666
Membrane	1024	831
Membrane-enclosed lumen	173	153
Nucleoid	1	1
Organelle	1250	1099
Synapse	124	79
Virion	22	20

3.4 The *C. floridanus* immunome

In total, 474 genes were identified as the constituent of the *C. floridanus* immunome based on an exhaustive analysis of literature, databases, and orthology prediction.

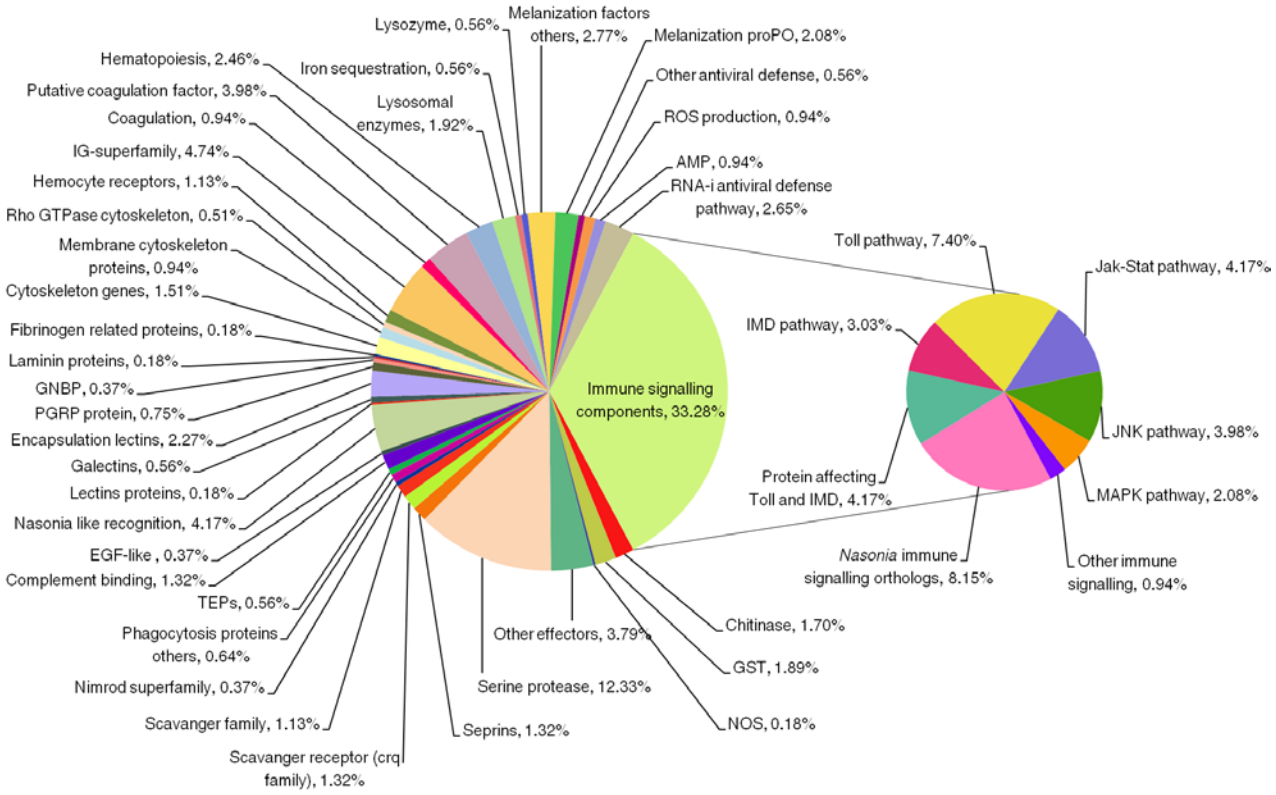


Figure 6. Functions of immune-related genes of *C. floridanus*. The pie chart shows the percentage of immune genes (474 genes in total) in each functional sub-category relative to the entire set of immune genes. The category 'immune signalling components' is further divided into different pathways [Image source: Gupta et al., 2015].

Immune genes from *D. melanogaster*, *A. mellifera*, and *N. vitripennis* were used as the reference to identify the immune genes in *C. floridanus* based on the concept that evolutionarily related genes which are the product of speciation events perform similar function. The reported 474 immune-related genes of *C. floridanus* were classified based on literature, GO annotation, and extensive manual curations. The genes were enriched in several

categories including 33.28 % of immune signalling components. Other important classified categories were microbial recognition genes, AMPs, phagocytosis, melanisation, encapsulation, cytoskeleton immune proteins, antiviral defence, coagulation, hematopoiesis and other immune responses (Fig. 6).

3.5 Comparison of the immune gene repertoire of *C. floridanus*

In terms of presented immune repository, including the alternative isoforms, 474 immune genes encode 510 proteins among which 307 proteins shares orthologs with *N. vitripennis*, 271 shares orthologs with *D. melanogaster*, and 221 with *A. mellifera*. Because these genes involve in multiple immune responses in different insects, they are likely to be important mediators/modifiers of the innate immune response. All the four species share 65 immune protein orthologs which mostly include the proteins involved in core immune signalling pathways, PRRs or serine proteases (Fig. 7, Appendix I).

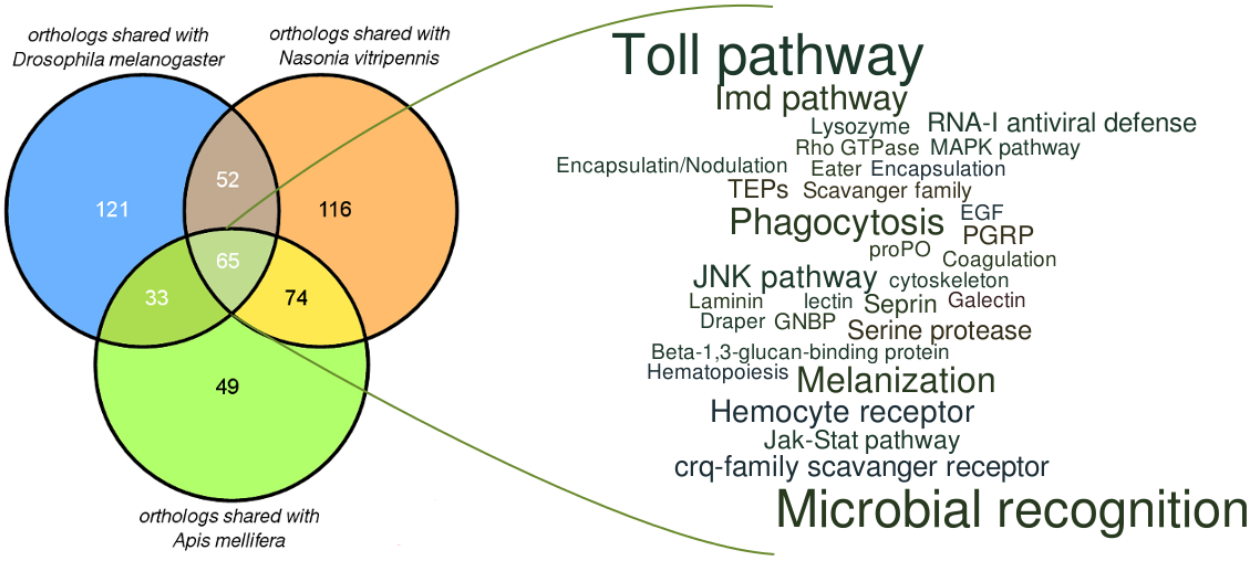


Figure 7. Immune gene repertoire of *C. floridanus* shared with different insect species. The Venn diagram shows the number of immune protein orthologs (including alternative splice isoforms) of *C. floridanus* shared with *D. melanogaster*, *N. vitripennis* and *A. mellifera*. Annotation of the 65 immune proteins shared by all three species is shown as word cloud where the size reflects the abundance of conserved protein function/immune annotation.

In agreement with the previous findings (Sackton et al., 2013) signalling pathways seem to be more conserved in comparison to effector molecules that show higher levels of taxon specificity. Moreover, the proteins implicated in microbial recognition were also identified to be highly enriched in the 65 conserved immune proteins.

3.6 Comparative genomics of ants, *A. mellifera, N. vitripennis* and *D. melanogaster*

Conservancy of *C. floridanus* proteins were analysed by comparing the re-annotated proteome with seven other sequenced ant species as well as *A. mellifera*, *N. vitripennis* and *D. melanogaster*. MCL clustering resulted in 18,763 groups from 188,092 protein sequences. All the eight ant species cluster into 6620 ortholog groups. Looking at the species distribution it was observed that 4,818 groups were shared by all the species analysed while 5,797 groups were shared by all hymenopterans. Altogether, 79 groups represent genes that are conserved exclusively among the eight ant species (Gupta et al., 2015). The 18,763 ortholog cluster were parsed with Perl scripts to reveal the presence of orthologs of *C. floridanus* immune proteins in selected species. Most of the proteins of *C. floridanus* involved in the humoral immunity signalling pathways were found to be mostly conserved in other ants and *A. mellifera*, however the some ortholog groups did not cluster *N. vitripennis* proteins, for instance, Traf6 cluster lack the ortholog of *N. vitripennis* protein (Gupta et al., 2015).

3.7 Antimicrobial peptides of *C. floridanus* and other hymenoptera

The previous analysis of the *C. floridanus* genome revealed a relatively low number of known AMPs including two defensins, a hymenoptaecin, a tachystatin-like and a crustin-like peptide (Ratzka et al., 2012; Tian et al., 2010; Zhang and Zhu, 2012). The apparently low number of AMPs may be compensated by the astonishing gene structure of hymenoptaecin which in the ants is encoded as a huge precursor protein with several repeated hymenoptaecin domains. In addition, the hymenoptaecin gene is among the most strongly induced genes after immune challenge. Most of the other ants also encodes these AMPs (Zhang and Zhu, 2012), however, the distribution pattern of AMPs is in general quite complex (Table 5), for instance, several of the ants including *C. floridanus* lack a gene encoding abaecin, while other ants encode this AMP. Thus, similar to the previously described gain, loss and duplication of defensin genes (Ratzka et al., 2012), there is quite an extensive variability in the presence and number of AMP genes in the ant genomes (Table 5). Much alike in the honeybee, the number of predicted or confirmed AMPs appears to be relatively low in ants as compared to the solitary wasp *N. vitripennis* for which 44 AMPs were described (Tian et al., 2010). In addition, ants produce a range of antimicrobial secretions that may be used to reduce pathogen pressure externally before an infection of the body occurs.

Table 5. Antimicrobial peptides (AMPs) of *C. floridanus* and other hymenoptera.

AMPs	*C. floridanus*	*H. saltator*	*L. humile*	*P. barbatus*	*A. cephalotes*	*S. invicta*	*A. echinatior*	*C. biroi*	*A. mellifera*	*N. vitripennis*
Hymenoptaecin[1)]	1	2	1	1	1	1	3	1	1	2
Defensin	2	2	1	5	1	2	1	1	3	5
Tachystatin-like	2	2	3	2	1	3	2	3	1	3
Crustin-like[2)*]	2	1	1	1	1	1	1	-	-	1
Abaecin	-	1	-	1	1	-	1	1	1	-
Melittin	-	-	-	-	-	-	-	-	1	-
Apisimin	-	-	-	-	-	-	-	-	1	-
Apidaecin	-	-	-	-	-	-	-	-	5	-
Navitripenicin	-	-	-	-	-	-	-	-	-	4
Nasonin[3)]	-	-	-	-	-	-	-	-	-	14
Nabaecin[4)]	-	-	-	-	-	-	-		-	4
Glynavici[3)]	-	-	-	-	-	-	-	-	-	7
Hisnavicin	-	-	-	-	-	-	-	-	-	5
Nahelixin	-	-	-	-	-	-	-	-	-	1

(1) Please note that here the number of genes present in the various species is indicated. In the ants the hymenoptaecin genes encode huge multipeptide precursor proteins which may give rise to several mature peptides (7 in the case of *C. floridanus*).
(2) One crustin-like AMP from Zhang and Zhu, 2012 and other identified from data in this thesis. *Modified, *C. floridanus* Crustin-like AMP from (Gupta et al., 2015).
(3) Adopted from Sackton et al. 2013.
(4) Nabaecin of *N. vitripennis* is considered as a member of the abaecin family. However, nabaecins belong to different orthologous cluster therefore, nabaecins and abaecins are separated here.

3.8 Prophenoloxidase, serine proteases and serpins

Several immune defence reactions in insects such as phagocytosis, melanisation and nodulation depend on phenoloxidase (PO) activity (Eleftherianos and Revenis, 2011; Lu et al., 2014; Sideri et al., 2008). During the melanisation process toxic intermediates such as reactive oxygen species (ROS) may kill microbial invaders directly. Since phenoloxidase activity can also harm insect cells, the enzyme is synthesised as an inactive precursor (Pro-PO). Pro-PO activation involves microbial recognition by PRRs and proteolytic cascades involving terminal serine proteases that finally cleave Pro-PO to its active form (Cerenius and Söderhäll, 2004; Lu et al., 2014). Serpins negatively control the activity of PO and help to avoid overshooting melanisation and dangerous ROS production (Tang, 2009). Phenoloxidases are related to arylphorins, hemocyanins and hexamerins (Hughes, 1999).

Using query sequences from four insect species, BlastP searches resulted in eight significant hits. The first hit corresponds to the *C. floridanus* prophenoloxidase (Cflo_N_g1918t1), while the other hits are distributed among hemocyanin, arylphorin and hexamerin sequences. Thus, a single prophenoloxidase gene appears to be present in the *C. floridanus* genome.

Additionally, 34 serine proteases and 10 serine protease inhibitors in *C. floridanus* (Appendix II) were annotated using the AutoFACT tool (Koski et al., 2005). Among these are five putative immune related serine proteases and four serine protease inhibitors including one serpin that showed differential expression profiles after immune challenge.

3.9 Chitinases, glutathione-S-transferases (GSTs) and nitric oxide synthase (NOS)

Enzymes with chitinase activity play an important role in the immune defence of insects by catalysing the breakdown of chitin, a linear polymer found in fungal pathogen cell walls consisting of β-1-4 linked N-acetylglucosamine (Aronstein et al., 2010). Four conserved motifs (KXXXXXGGW, FDGXDLDWEYP, MXYDXXG and GXXXWXXDXD, where X is a non-specified amino acid) have been reported in catalytic domains of insect chitinases (De la Vega et al., 1998; Kramer and Muthukrishnan, 1997). Five prototypic chitinases from different insect species served as a standard for detection (Choi et al., 1997; Li et al., 2007; Shen and Jacobs-Lorena, 1997; Zhu et al., 2004). In *C. floridanus* 13 putative chitinases were found containing a variable number of the four conserved sequence motifs (Appendix III). Only two predicted proteins are endowed with all four chitinase signature motifs, while in five sequences no such motif was detected. Overall, there is not much variation in the number of chitinase genes encoded by the different insect species compared here (Appendix IV).

Glutathione S-transferases (GSTs) comprise a diverse family of dimeric enzymes that have attracted attention in insects because of their involvement in the defence towards insecticides (Chelvanayagam et al., 2001). Cytosolic GSTs in insects have been assigned to six classes including delta, epsilon, omega, sigma, theta and zeta (Friedman, 2011; Ranson et al., 2001), and among them the delta and epsilon classes represent over 65 % of the total GST expansions. A recent phylogenetic study of insect GSTs suggested the evolution of the epsilon class from the delta class (Aronstein et al., 2010). Several *C. floridanus* GSTs were classified into different classes on the basis of their sequence similarities and phylogenetic relationships with other insect species. Based on these approaches, nine out of ten identified *C. floridanus* GSTs were assigned to five different classes, including three in omega, three in sigma, one in

each of delta, theta and zeta and one unclassified GST (Appendix V). The absence of epsilon class GSTs in *C. floridanus* is in line with their absence in other hymenoptera (Foley and O'Farrell, 2003). With the exception of *D. melanogaster* encoding 20 GSTs, there are only minor differences in the number of GSTs encoded by the other insects (between eight and eleven GSTs) (Appendix IV).

NOS belongs to the family of enzymes which form nitric oxide (NO) from L-arginine and makes important contributions to the IMD pathway in activation of Relish, since NOS activity is required for a robust innate immune response to Gram-negative bacteria in *D. melanogaster* (Davies et al., 2012; Eleftherianos et al., 2014). In *C. floridanus* only a single gene (Cflo_N_g5430t1) codes for a protein that matched all criteria of a NOS. With the exception of *A. cephalotes* and *N. vitripennis* each encoding two NOS, the other ants and *A. mellifera* code for a single NOS (Appendix IV).

3.10 Signal transduction via major immune signalling pathways

Exploiting the in-depth knowledge and functional assays of *D. melanogaster* major immune signal transduction pathways (Toll, Jak-Stat, IMD, and JNK) (Hoffmann, 2003; Kounatidis and Ligoxygakis, 2012), the humoral immunity signalling pathways of *C. floridanus* were derived which showed high conservancy with *D. melanogaster* due to the presence of orthologs of most of the signalling components.

3.10.1 Toll signalling pathway

The Toll pathway is required for the host response against fungal and most of the Gram-positive bacterial infections. The pathway not only regulates the antimicrobial response but is also required for proper haemocyte proliferation (Lemaitre and Hoffmann, 2007; Qiu et al., 1998), hence it leads to a coordinated immune response that comprises both cellular and humoral immunity (Valanne et al., 2011). Besides the PRR PGRP-SD and DIF, all the component of *C. floridanus* Toll signalling pathway was found to be highly conserved in terms of the presence of homologs in *Drosophila* (Fig. 8). The absence of PGRP-SD indicates the Toll pathway in *C. floridanus* is not activated via this signalling after infection with Gram-negative bacteria. Altogether, three PRRs were annotated that may trigger the Toll cascade in *C. floridanus* which include PGRP-SA (Cflo_N_g8526t1) and two proteins annotated as beta-1,3-glucan binding proteins (Cflo_N_g15215t1 and Cflo_N_g5742t1) with high homologies to both GNBP1 and GNBP3 of *D. melanogaster*.

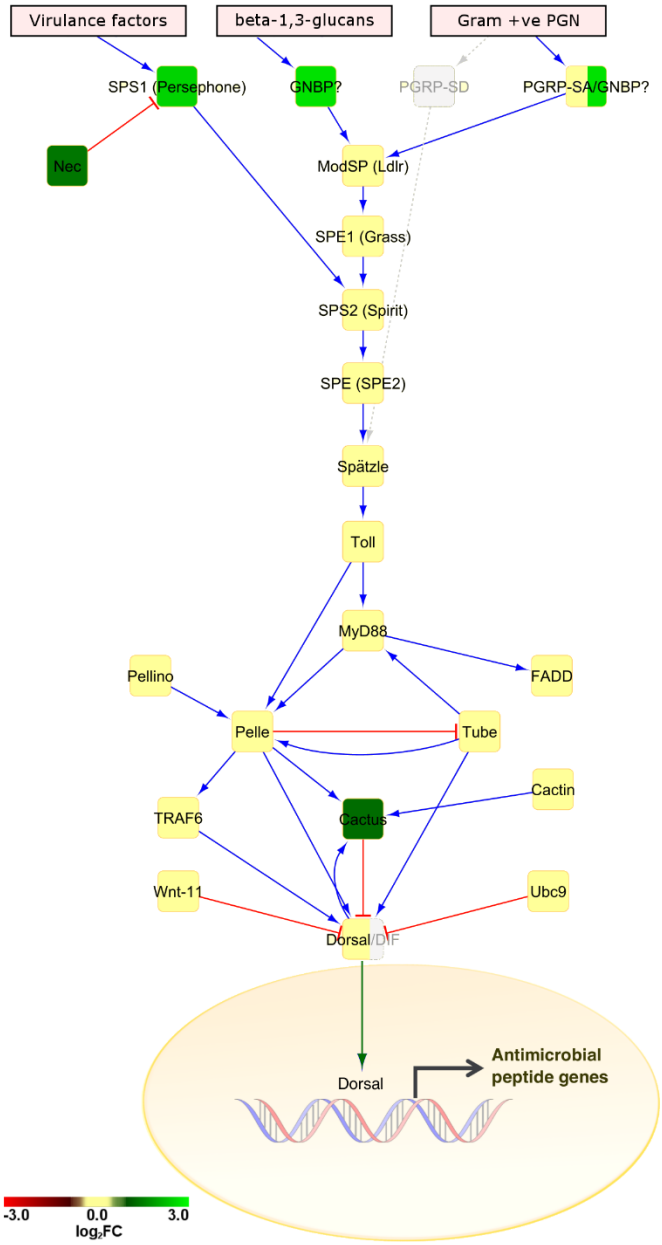


Figure 8: The Toll signalling pathway of *C. floridanus*. All identified signalling components are mapped on the comprehensive immune network of *D. melanogaster*. The names of the factors correspond to the *Drosophila* designations. Connectivity among nodes is based either on positive attribute (blue arrow) or negative attribute (red arrow). Missing components are shown in grey colour. Nuclear translocation is shown by a green arrow. Factors significantly upregulated on the transcriptional level upon immune-challenge are shown by green boxes.

The induced Toll signalling cascade finally stimulates the nuclear translocation of the nuclear factor kappa-B (NF-κB) factors Dorsal and Dorsal-related immunity factor (DIF). In agreement with the observation that DIF is a highly derived branch found in brachyceran flies but absent in *A. mellifera* (Evans et al., 2006) and other insects, no orthologous of DIF was found in the *C. floridanus*. Therefore, in *C. floridanus* Dorsal appears to be required for the induction of the transcription of AMPs during the Toll mediated immune response.

3.10.2 IMD and JNK signalling pathways

In insects IMD pathway is activated predominantly by infection of Gram-negative bacteria. Diaminopimelic acid-type (DAP-type) peptidoglycan (PGN) from bacteria is recognised by single signal-transducing PRR i.e., PGRP-LC receptor (Cflo_N_g10272t1) in *C. floridanus* to trigger the IMD signalling, while the organism lacks PGRP-LE (Fig. 9).

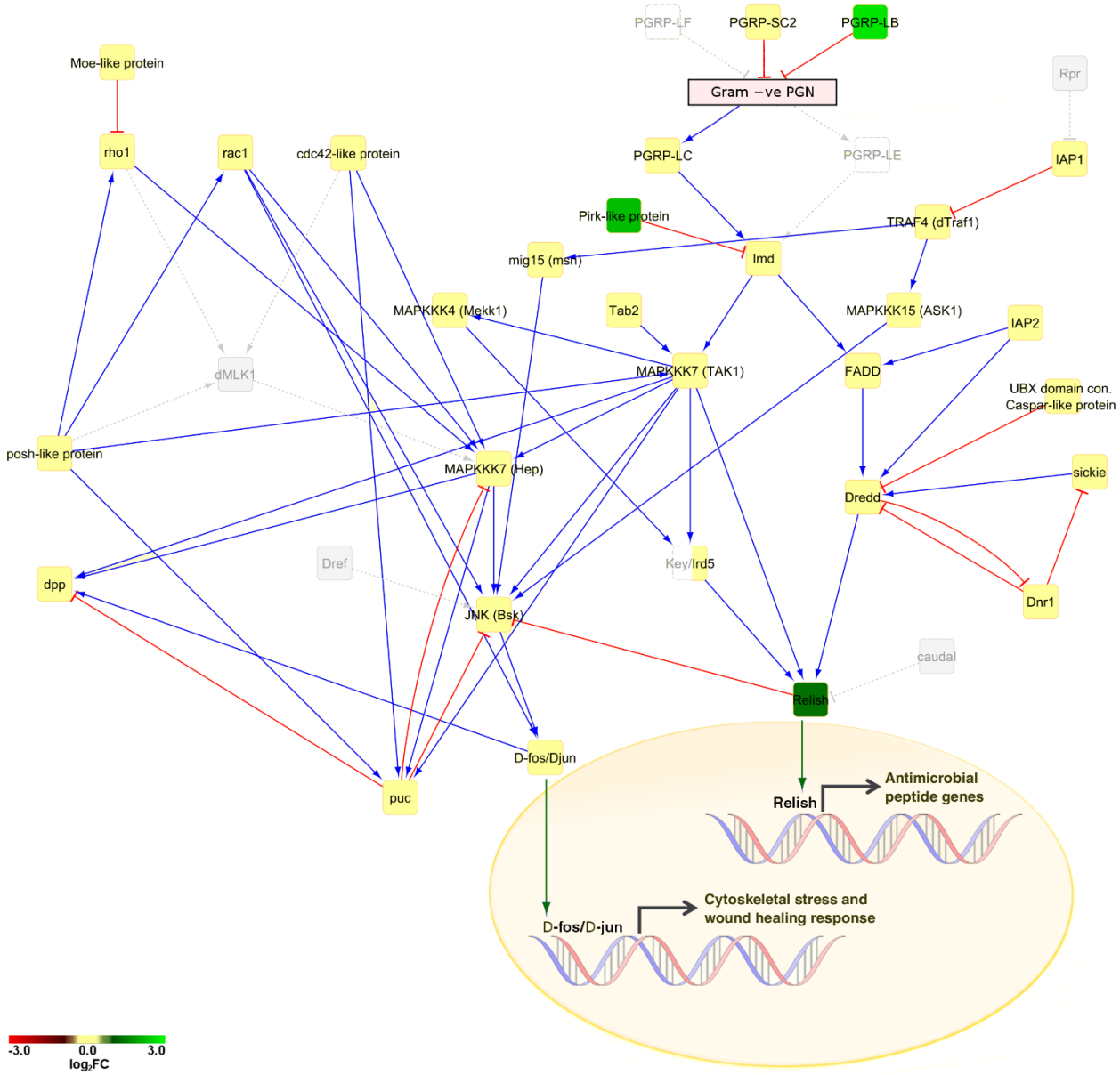


Figure 9: The IMD and JNK pathways of *C. floridanus*. All identified signalling components are mapped on the comprehensive immune network of *D. melanogaster*. The names of the factors correspond to the *Drosophila* designations. Connectivity among nodes is based either on positive attribute (blue arrow) or negative attribute (red arrow). Missing components are shown in grey colour. Nuclear translocation is shown by a green arrow. Factors significantly upregulated on the transcriptional level upon immune-challenge are shown by green boxes.

Besides one PGRP-LE receptor and one subunit (Kenny) of *Drosophila* IKK complex, homologs of all the other core components of IMD pathway were identified in *C. floridanus* (Fig. 9). Upon activation the IKK complex activates the NF-κB-like transcription factor Relish by phosphorylation (Hasdemir et al., 2009). Additionally, the comparative analysis revealed the absence of Kenny subunit in *A. mellifera*, *N. vitripennis* and other ant species, suggesting a common character of the IKK complex in hymenoptera.

The IMD pathway also leads to TAK1 (transforming growth factor β-activated kinase 1) mediated activation of JNK signalling cascade (Silverman et al., 2003). Orthology analysis suggested the core components of the JNK signalling pathway are conserved in *D. melanogaster* and *C. floridanus*. The homologs of *Drosophila* JNK component, Mlk1 and Dref were not identified in *C. floridanus* JNK cascade.

3.10.3 Jak-Stat signalling pathway

The presence of corresponding homologs of the components of *Drosophila* Jak-Stat signalling pathway in *C. floridanus* were analysed. It was observed that the *C. floridanus* immune system is enriched with all the core components of Jak-Stat pathway (Fig. 10) except the ligand homologs which suggested the existence of Jak-Stat pathway in *C. floridanus*, triggered by unknown ligands. In *Drosophila* the downstream effector molecules of Jak-Stat pathway are thioester-containing proteins (TEPs) and turandots (tots) proteins (Agaisse et al., 2003; Lagueux et al., 2000) however, similar to *A. mellifera* no homologs of Turandot proteins were identified in *C. floridanus*, but several phagocytosis promoting molecule TEPs were found.

Most TEPs share the common CGEQ motif defining the thioester site, which allows the formation of a covalent bond to microbial surfaces (Bou Aoun et al., 2010) but several TEPs in insects lack the thioester motif (Blandin and Levashina, 2004). Similarity search revealed the presence of three TEP genes (TEP1: Cflo_N_g7345t1, TEP2: Cflo_N_g4492t1, and TEP3: Cflo_N_g9745t1) in *C. floridanus*, among them TEP1 and TEP2 consist the CGEQ motif. For the gene encoding TEP3, two alternative transcripts were found (Cflo_N_g9745t1 and Cflo_N_g9745t2). Interestingly, only the CGEQ motif containing TEP1 of *C. floridanus* was found to be significantly up-regulated upon immune challenge.

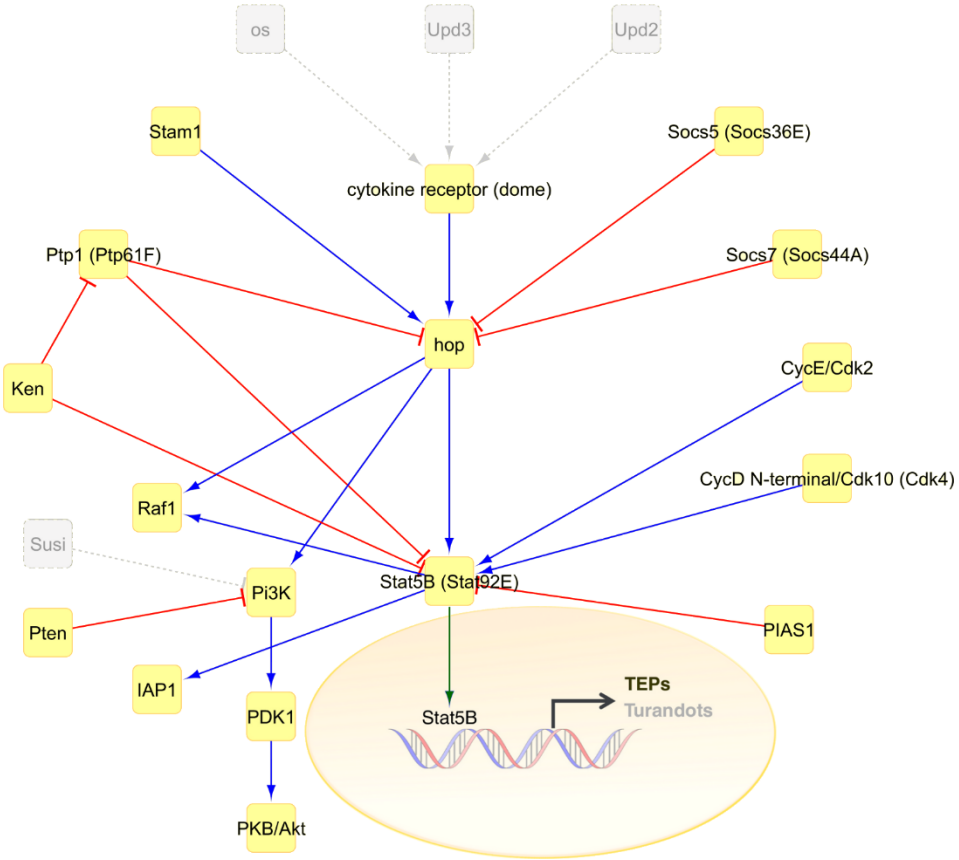


Figure 10: The Jak-Stat pathway of *C. floridanus*. All identified signalling components are mapped on the comprehensive immune network of *D. melanogaster*. The names of the factors correspond to the *Drosophila* designations. Connectivity among nodes is based either on positive attribute (blue arrow) or negative attribute (red arrow). Missing components are shown in grey colour. Nuclear translocation is shown by a green arrow.

3.11 Identification of differentially expressed genes (DEGs) after immune challenge

The DEGs were analysed using Illumina sequencing by comparing transcript abundance of bacterial-challenged *C. floridanus* with the untreated *C. floridanus* group. Cuffdiff (q-value < 0.05 and fold change ≤ 2.0) and DESeq (p-value adjusted < 0.05 and fold change ≤ 2.0) analysis of the *C. floridanus* samples revealed a total of 257 transcripts that showed significant changes in expression in response to bacterial challenge.

Among the DEGs, ~56 % were up-regulated, and ~ 44 % were down-regulated. Moreover, among these transcripts, ~ 20 % of transcripts were identified to code for known immune related proteins. To show at the same time amount of change and statistical significance, Volcano plots summarise the results (Fig. 11). All the DEGs including their annotation and log fold change expression value is listed in Appendix VI. Genes up-regulated after immune challenge encode well known immune related genes including those that encode

PRRs and serine proteases (e.g. Snake and Stubble-like), proteins involved in signalling and transcription (e.g. nuclear factor NF-κB p110 subunit (Relish), NF-κB inhibitor (Cactus), as well as stress-related proteins such as cytochromes P450.

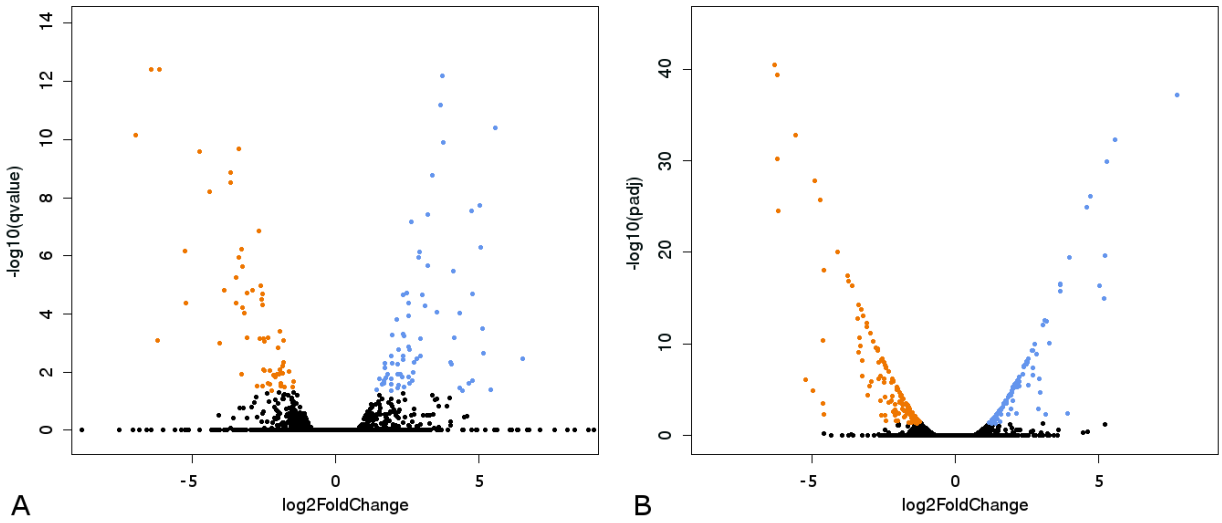


Figure 11. Gene expression changes after immune challenge. Volcano plots show the statistical significance of the difference in expression observed (p-value from a t-test, q-value in case of Cuffdiff and adjusted p-value in case of DESeq; log10 scale). The x-axis indicates the differential expression profiles, plotting the fold-induction ratios in a log-2 scale during immune challenge. The list of significantly differentially expressed protein coding genes can be accessed from (Gupta et al., 2015). (A) Volcano plot from Cuffdiff data. Up-regulated genes (q-value < 0.05 and log2FoldChange ≥1) are shown as blue dots and the down-regulated genes (q-value < 0.05 and log2FoldChange ≤ −1) are shown as orange dots. The top three protein coding genes most up-regulated are Cflo_N_g6748 (hypothetical), Cflo_N_g2827 (voltage-dependent calcium channel type A subunit alpha-1) and Cflo_N_g12631 (serine proteinase Stubble). The three most down-regulated genes include Cflo_N_g2215 (putative chitin binding peritrophin-a domain containing), Cflo_N_g1319 (serine protease inhibitor 3) and Cflo_N_g4308 (lipase member H). (B) Volcano plot from DESeq data. Up-regulated genes (adjusted p-value < 0.05 and log2FoldChange ≥1) are shown by blue dots and the down-regulated genes (adjusted p-value < 0.05 and log2FoldChange≤ −1) are shown by orange dots. The top three protein coding genes most up-regulated are Cflo_N_g6748 (hypothetical), Cflo_N_g5222 (hypothetical) and Cflo_N_g531 (aminopeptidase N). The three top down-regulated genes include Cflo_N_g2215 (putative chitin binding peritrophin-a domain containing), Cflo_N_g1319 (serine protease inhibitor 3) and Cflo_N_g907 (chymotrypsin-1) [Image source: Gupta et al., 2015].

Nine DEGs encoding putative immune-related proteins of *C. floridanus* do not have any homologs in other insects. EuKaryotic Orthologous Groups (KOG) annotations of these proteins reveal features of some of these proteins including the presence of signal peptides, a chemosensory domain, and a DNA-binding domain (listed in Appendix VII). Moreover, in a recent study Hamilton and co-workers reported Cathepsin D as a protein that contributes to social immunity in *Camponotus pennsylvanicus* (Hamilton et al., 2010). *C. floridanus* also

encodes an ortholog of Cathepsin D (Cflo_N_g9172t1) and it will be interesting to investigate a general role of this protein in social immunity of *C. floridanus* in the future.

Among the DEGs, a hypothetical protein (Cflo_N_g6748t1) encoding gene was found to be highly up-regulated. The differential expression of gene encoding this protein was also confirmed by qRT-PCR. Both in immune-challenged larvae and workers, Cflo_N_g6748t1 showed high induction (Appendix VIII). This protein consists a domain of unknown function (DUF2236) however, several hymenopterans consist the homologs of this protein, including the ants (*H. saltator, S. invicta, A. cephalotes, A. echinatior, C. biroi* and *Lasius niger*), bees (*A. mellifera, Habropoda laboriosa, Melipona quadrifasciata*), wasps (*N. vitripennis* and *Fopius arisanus*) which indicates the protein might have some conserved function. The extreme expression pattern of this protein in *C. floridanus* during infection suggest that it might play an important role in the immune defence of larvae and might merit future attention.

The validation of DEGs was performed by qPCR to determine the expression levels of 15 up-regulated and 12 down-regulated genes. The immune challenged larvae and adult animals were treated separately instead of using pooled developmental stages RNA which was used for generation of the Illumina sequencing data. The qRT-PCR analysis confirmed the Illumina data showing that the selected genes are regulated in the same direction in both larvae and workers, or at least in one of the two developmental stages (Fig. 12).

The genes with the corresponding fold changes and significance value is listed in Appendix VIII. Notably, it was observed that change in expression of several genes after immune-challenge differs substantially between larvae and workers. However, this is in agreement with published literature on other insects which indicated that the immune response of larvae might differ from that of adults in holometabolous insects including workers in social insects (Fellous and Lazzaro, 2011; Randolt et al., 2008; Rosengaus et al., 2013).

Overall, the induction of gene expression by immune challenge with bacteria appeared to be much stronger in larvae than in workers which was further in agreement with MS analysis of *C. floridanus* haemolymph (explained later in this thesis). The difference in expression of larvae and adults were observed in several genes such as MPI (*metalloproteinase inhibitor*), Socs2 (*suppressor of cytokine signalling 2*), Cact1 (*cactus*), Transf (*transferrin*), PHR (*parathyroid peptide receptor*), ester (*esterase FE4*), PGRP-LB, PGRP-SA, Hp67112 (hypothetical protein, Cflo_N_g6748t1), *hymenoptaecin*, and *thioester-*

containing protein 1 (TEP1) was strongly induced in immune-challenged larvae, but only weakly or not at all in workers (Fig. 12; Appendix VIII).

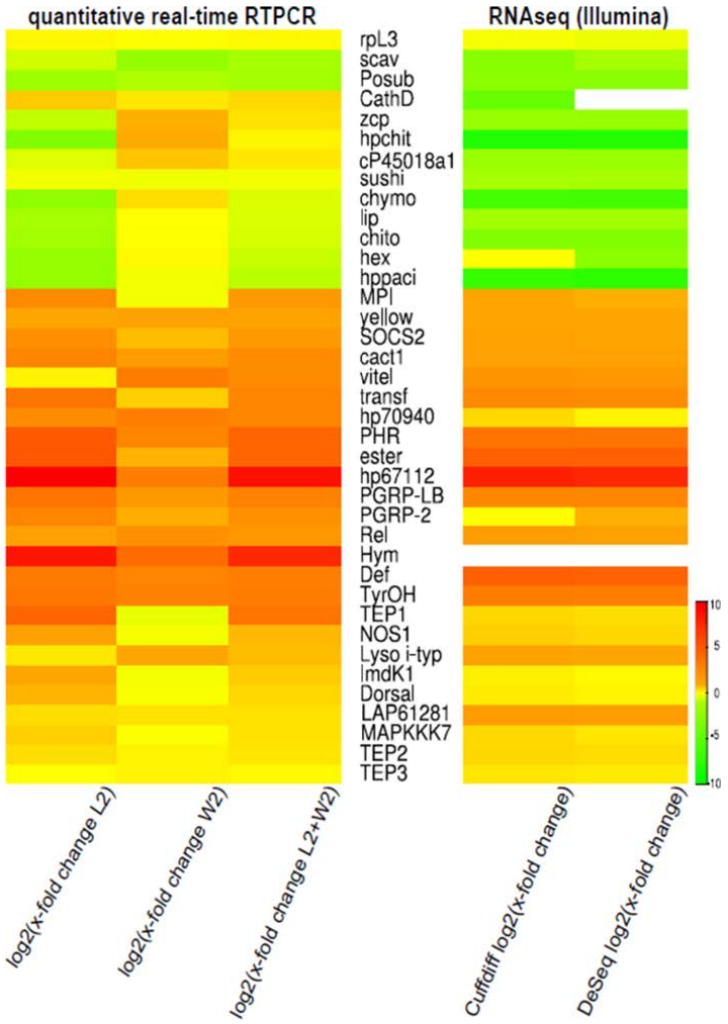


Figure 12. Comparison of expression of 37 selected genes based on the analysis of the Illumina sequencing data using Cuffdiff and DESeq and the corresponding qRT-PCR data. The heat map visualises the expression of 37 regulated immune genes 12 h after pricking of larvae and workers with a 1:1 mix of Gram-negative and Gram-positive bacteria in case of Illumina sequencing (Cuffdiff log2FC / DESeq log2FC). The corresponding qRT-PCR analysis was performed separately in larvae and workers revealing a stage-specific gene regulation after immune challenge (log2(x-fold change L2 + W2)). Up and down-regulation are colour coded as given in the key highlighting common directions of gene regulation in the different data sets [Image source: Gupta et al., 2015; prepared by Maria Kupper].

The orthologous comparison of *C. floridanus* differentially expressed immune genes with immune-stimulated *A. mellifera* DEGs (Richard et al., 2012) revealed the presence of six common genes which encodes the immune signalling pathway genes (Cactus, Relish, Stubble and Seprin B1), Tyrosine 3-monooxygenase (involved in melanisation and proPO pathway in *Manduca sexta* (Gorman et al., 2007)), and a protein NPC2 homolog (involved in microbial recognition in *D. melanogaster* (Shi et al., 2012)).

Additionally, similar to several other insects the down-regulation of multiple genes that encode proteins involved in digestion (e.g. Chymotrypsin, Lipase) and storage (e.g. Hexamerin) after immune challenge of *C. floridanus* (Appendix VIII) indicates that during infection insects seem to temporarily shut down digestion and synthesis of non-essential proteins in order to use resources for costly defence reactions (Aguilar et al., 2005; Ayres and Schneider, 2009; Lourenço et al., 2009; Meng et al., 2008).

3.12 Functional annotation of DEGs

Using Blast2GO platform, GO terms for DEGs were obtained and Fisher's exact test was applied to identify GO terms that were overrepresented at p-value ≤ 0.05 relative to the whole *C. floridanus* proteome. By relying on semantic similarity measures, the GO filtering clusters the related GO terms together in the enriched sets. The less specific GO terms in the clusters are represented by one GO term while prioritizing the more statistically significant term in the same cluster. In total, 44 GO terms overrepresented in the DEGs include several implicated important categories overrepresented categories, for instance, oxidoreductase activity (p-value 4.08E-06), catalytic activity (p-value 5.73E-04), hydrolase activity (p-value 1.01E-03), peptidase activity (p-value 2.32E-02), regulation of homeostatic process (p-value 3.27E-02) and peptidoglycan metabolic process (p-value 4.86E-02). Table 6 lists the GO terms that were overrepresented in the set of DEGs after filtering for the most specific GO terms.

Table 6. GO terms significantly enriched in the set of genes that are differentially expressed in *C. floridanus* upon immune challenge. Category F represents molecular function, category P represents biological function and category C represents cellular compartments.

GO-ID	Term description	Category	P-value	Significance
GO:0016491	oxidoreductase activity	F	4.08E-06	***
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	F	4.00E-04	***
GO:0003824	catalytic activity	F	5.73E-04	***
GO:0016787	hydrolase activity	F	1.01E-03	**
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	F	2.70E-03	**
GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	F	2.95E-03	**
GO:0016798	hydrolase activity, acting on glycosyl bonds	F	5.19E-03	**
GO:0005506	iron ion binding	F	8.54E-03	**

GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	F	8.54E-03	**
GO:0004497	monooxygenase activity	F	9.49E-03	**
GO:0004459	L-lactate dehydrogenase activity	F	1.65E-02	*
GO:0004457	lactate dehydrogenase activity	F	1.65E-02	*
GO:0042132	fructose 1,6-bisphosphate 1-phosphatase activity	F	1.65E-02	*
GO:0008271	secondary active sulfate transmembrane transporter activity	F	1.65E-02	*
GO:0016712	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	F	1.65E-02	*
GO:0016831	carboxy-lyase activity	F	2.15E-02	*
GO:0008233	peptidase activity	F	2.32E-02	*
GO:0004622	lysophospholipase activity	F	3.27E-02	*
GO:0004511	tyrosine 3-monooxygenase activity	F	3.27E-02	*
GO:0008395	steroid hydroxylase activity	F	3.27E-02	*
GO:1901682	sulfur compound transmembrane transporter activity	F	3.27E-02	*
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	F	3.52E-02	*
GO:0008745	N-acetylmuramoyl-L-alanine amidase activity	F	4.86E-02	*
GO:0009982	pseudouridine synthase activity	F	4.86E-02	*
GO:0008199	ferric iron binding	F	4.86E-02	*
GO:0009450	gamma-aminobutyric acid catabolic process	P	1.65E-02	*
GO:0006706	steroid catabolic process	P	1.65E-02	*
GO:0018214	protein carboxylation	P	1.65E-02	*
GO:0008272	sulfate transport	P	1.65E-02	*
GO:0017187	peptidyl-glutamic acid carboxylation	P	1.65E-02	*
GO:0009448	gamma-aminobutyric acid metabolic process	P	3.27E-02	*
GO:0032844	regulation of homeostatic process	P	3.27E-02	*
GO:0051899	membrane depolarization	P	3.27E-02	*
GO:0016264	gap junction assembly	P	3.27E-02	*
GO:0010644	cell communication by electrical coupling	P	3.27E-02	*
GO:0009713	catechol-containing compound biosynthetic process	P	3.27E-02	*
GO:0072348	sulfur compound transport	P	3.27E-02	*
GO:0006826	iron ion transport	P	4.86E-02	*
GO:0000270	peptidoglycan metabolic process	P	4.86E-02	*
GO:0006013	mannose metabolic process	P	4.86E-02	*
GO:0010496	intercellular transport	P	4.86E-02	*
GO:0005576	extracellular region	C	1.80E-02	*
GO:0031225	anchored component of membrane	C	4.86E-02	*
GO:0032156	septin cytoskeleton	C	4.86E-02	*

Symbols: p-value ≤ 0.001 = ***, p-value ≤ 0.01 = **, p-value ≤ 0.05 = *

3.13 Interactome of *C. floridanus*

Proteins interacting in one organism co-evolve such that their respective orthologs maintain the ability to interact in another organism (Sharan et al., 2005; Yu et al., 2004). The orthologous interacting proteins pair is termed as interolog of template interacting protein pair

present in other species. The preliminary interactome of *C. floridanus* was reconstructed using the interolog analysis against all the available high resolution experimental PPI maps for *D. melanogaster* and the PPI data from DIP database. The resulted interactome consists 6274 nodes and 51866 edges. The accuracy of interactome can be increased by adding the interacting domain information (Wojcik and Schachter, 2001; Zhang et al., 2009; Zhou et al., 2013). Out of 51866 interolog based *C. floridanus* PPIs 20544 interactions were identified to have at least one pair of interacting domains. The number of nodes in the network also reduced by this filtering step and resulted in 4589 nodes. Since the many interactions detected would not be plausible if both the interacting partners do not have any shared subcellular localization, the DDI supported network was pruned based on localization information of the interactors. In case of single localization only the interactions in the same cellular compartment were considered while in case of multiple localizations the interactions were considered true only if interactors have at least one similar cellular compartment annotation. This filtering step resulted in high confidence *C. floridanus* interactome consisting 3914 nodes and 13640 edges.

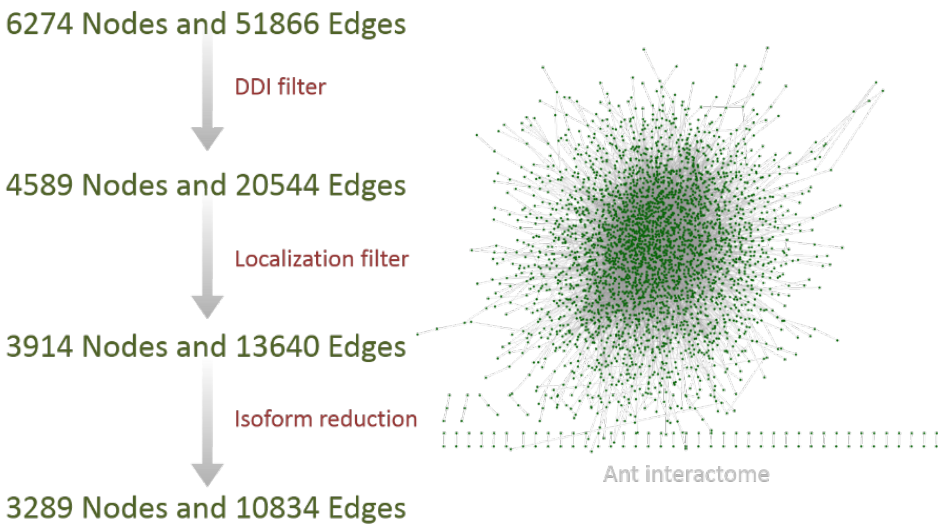


Figure 13. Schematic diagram depicting the filtering steps to reconstruct the draft interactome of *C. floridanus*.

Moreover, to avoid the additional complexity of the interactome the genes that encodes multiple isoforms were represented with single node if the alternative forms were found to be involved in similar interactions. The final interactome of *C. floridanus* (IC) consist 3289 nodes and 10834 edges and the reconstruction framework is depicted in a schematic overview in Fig. 13.

3.14 Network analysis of *C. floridanus* interactome

The topological properties analysis of an interactome provides a view of the network as a system and assists for the identification of the components that play a central role in network connectivity. The network diameter and the mean path length were found as 14 and 4.3, respectively, indicating the small-world topology. The connectivity (k, the number of links per node) distribution of the nodes in the reconstructed *C. floridanus* interactome was found as scale-free, as degrees of nodes in network are distributed approximating with the power-law model ($P(k) \sim k^{-\gamma}$) where P(k) is the probability of a node having a degree of k and $\gamma > 0$. The *C. floridanus* interactome showed a clustering coefficient of 0.094 and an average degree of 6.970. Overall, consistent with general characteristic of complex biological networks (Giot et al., 2003; Jeong et al., 2001; Maslov and Sneppen, 2002) the interactome indicated the presence of scale-free features with small-world topology.

3.15 Signalome of *C. floridanus*

Intracellular signalling is viewed as a set of intertwined pathways forming a single signalling network (Papin et al., 2005). Signal transduction pathways are involved in the control of various cellular processes, including cell growth, proliferation, differentiation and stress response in divergent animal phyla (Pires-daSilva and Sommer, 2003). Different resources were used to reconstruct the first genome scale signalling map of *C. floridanus* by mapping of ortholog pair of proteins involved in signalling interactions. Briefly, the concept of signalog was implemented which states two proteins are predicted to involve in signalling, if their ortholog pair in another organism also involves in signalling (Korcsmáros et al., 2011). The signalome of *C. floridanus* (SC) represented 2199 nodes connected by 4203 edges. The reconstructed SC consist 3001 stimulatory and 1202 inhibitory connections. The proteasome complex related proteins were identified as the top hubs in SC, which revels the complex can handle many diverse signals while the multifunctional serine/threonine-protein kinases were recognised as top bottlenecks. Fig. 14 shows the graphical representation of draft SC with all the stimulatory and inhibitory connections.

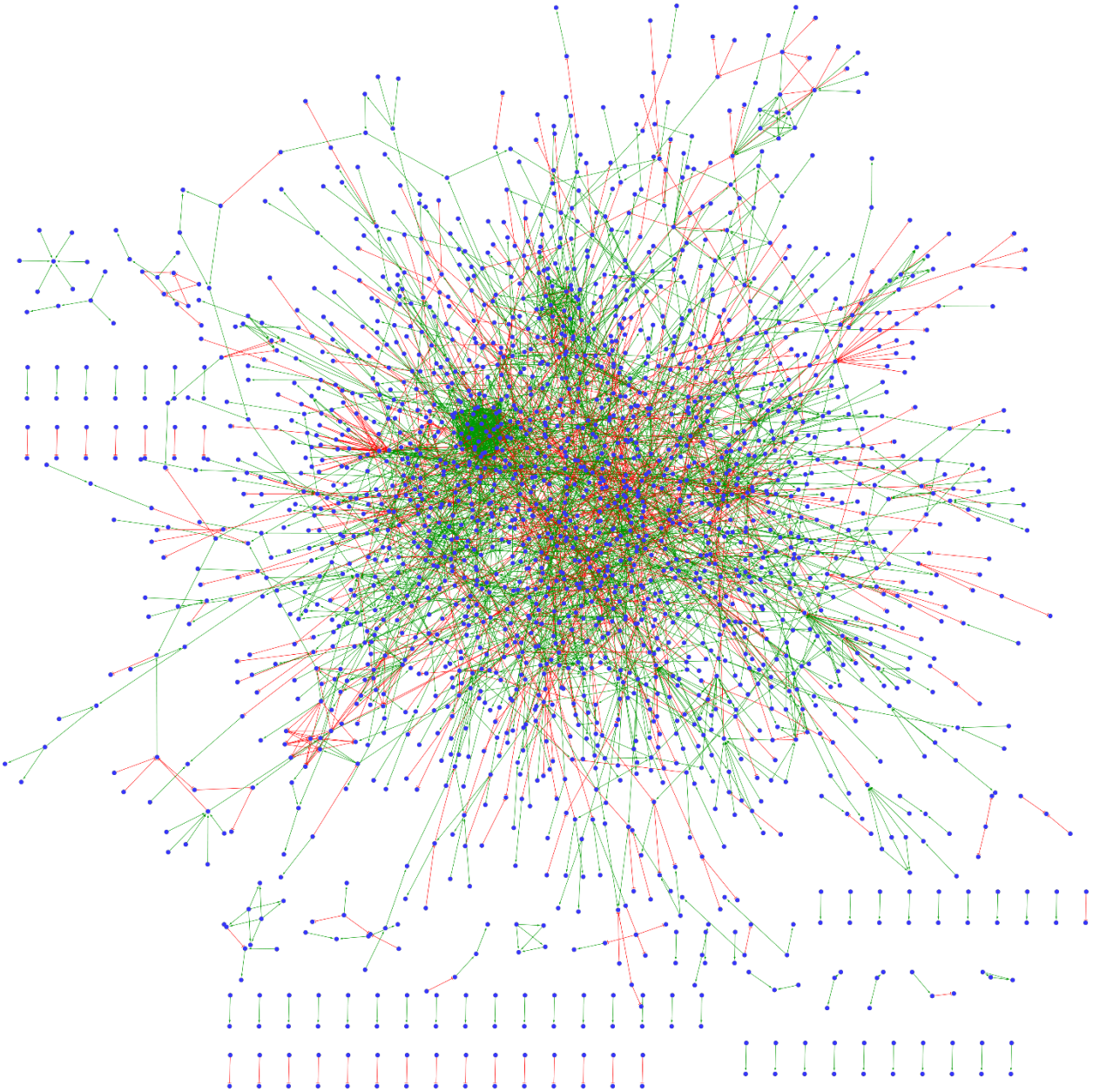


Figure 14. Overview of reconstructed *C. floridanus* signalome (see Appendix IX for annotation of top hubs and bottlenecks). Connectivity among nodes is based either on activation (→), inhibition (⊣). Edge description: red, inhibition; green, activation.

3.16 Integration of interactome and signalome

Integrating IC and SC data sets using uniform manual curation criteria can significantly contribute to a more precise assessment of the network and could be exploited to understand the underlying phenomenon of multiple biological process. Network analysis identifies 18 topologically important proteins that showed the expression changes in response to bacterial infection. Notch (Cflo_N_g6260), Relish (Cflo_N_g6082), and SgAbd-1 (Cflo_N_g7775) were determined as top three hubs while Relish, Notch and GNBP

(Cflo_N_g5742) were found as top three bottlenecks. The first order connections of the differentially expressed hubs bottlenecks are depicted in Fig. 15. Moreover, all the DEGs were mapped on IISC and the DEGs present in IISC were connected by shortest path algorithm.

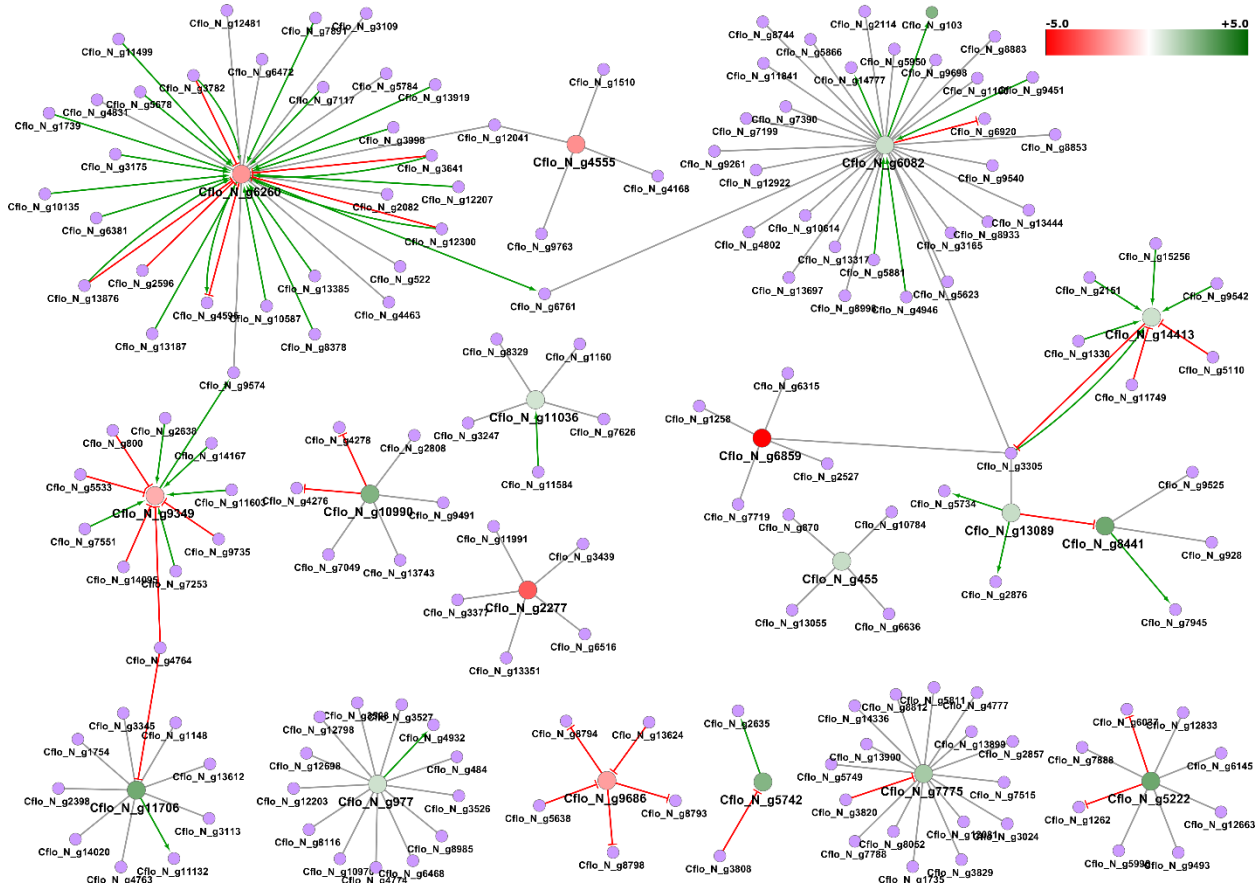


Figure 15. The primary interactors of the top 20 % of differentially expressed hubs and bottlenecks (big size nodes; see Appendix X for annotation of top hubs and bottlenecks). Node designation: red gradients, down-regulated genes; green gradients, up-regulated genes; violet, not differentially expressed. All nodes are denoted by re-annotation identifiers (see Appendix VI for annotation and expression value). Connectivity among nodes is based either on activation (→), inhibition (—|) or undirected interactions (—). Edge description: red, inhibition; green, activation; grey, no regulation.

Altogether, 72 DEGs were mapped on IISC and 68 DEGs were connected by shortest path and resulted in IISC subnetwork. The IISC subnetwork were further functionally annotated based on homology, GO terms, and literature. Several immune related protein clusters were annotated in the subnetwork (Fig. 16).

By integrating interactions in IISC with mRNA expression data the active subnetwork of the IISC were identified that shows the enriched connected region with differentially expressed genes. In the active module, the regulation of Lysozyme c-1 (Cflo_N_g4036) was

identified which could play an important role in antibacterial immunity (Kajla et al., 2010; Kajla et al., 2011).

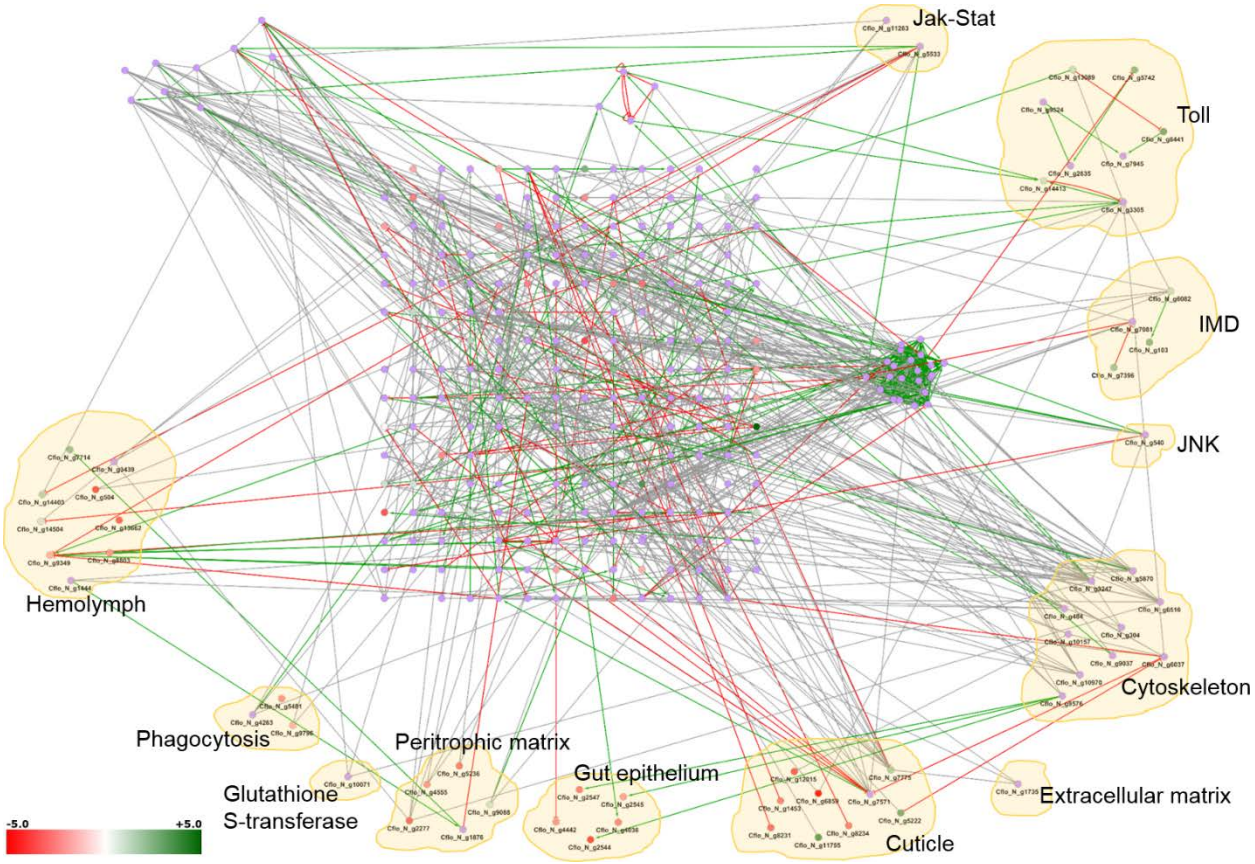


Figure 16. Subnetwork extracted from IISC network based of shortest path connecting the differentially expressed proteins. Node designation: red gradients, down-regulated genes; green gradients, up-regulated genes; violet, not differentially expressed. All nodes are denoted by re-annotation identifiers (see Appendix VI for annotation and expression value). Connectivity among nodes is based either on activation (→), inhibition (—|) or undirected interactions (—). Edge description: red, inhibition; green, activation; grey, no regulation.

3.17 Interaction of *C. floridanus* with pathogen *S. marcescens*

The analysis of *C. floridanus* immune system reveled the evidence of multilayered architecture. The host-pathogen PPIs resulted in 1127 putative interactions. Notably, not all the identified host-pathogen PPIs have relevance in bacterial pathogenesis but these are biologically viable interactions occurring across the host and pathogen. Not only the pathogenic factors but also the several surface proteins, membrane proteins, other secreted proteins and digestive enzymes of pathogen participate in the host interactions. Here for the sake of simplicity, these are mentioned as bacterial proteins in the upcoming text. Nevertheless, dividing the element of immune system in specific categories is somewhat arbitrary as some proteins could plausibly be assigned to multiple categories, for instance,

AMPs can be produced both in gut epithelium and fat-bodies, however, ant proteins involve in exciting interactions with pathogen and often placed here in suitable categories.

3.17.1 Interactions with peritrophic matrix proteins

The chitin-binding domain containing peritrophic matrix (PM) proteins can modulate the structure and porosity of the PM (Dinglasan et al., 2009). An interaction of chitin binding peritrophin-A domain containing acidic mammalian chitinase like PM protein (Chia; Cflo_N_g2277) of *Camponotus* with a transmembrane protein LOLE of *S. marcescens* was identified. Interestingly, the downregulation of chiA gene was observed in immune challenged *C. floridanus*.

3.17.2 Interactions with gut epithelium

The production of reactive oxygen species (ROS) through the NADPH dual oxidase (DUOX) by epithelial innate immune system constitutes the strong inducible insect defence. Several G-protein coupled receptors (GPCRs) contributes to DUOX-modulated gut physiology during gut-microbe interactions (Kim and Lee, 2014). Four GPCRs (Cflo_N_g11385, Cflo_N_g1673, Cflo_N_g14626, Cflo_N_g12263) were found in interaction with *Serratia* proteins, however further experimental studies are essential to illustrate the role of these interactions in DUOX stimulations. Several hubs for IC were annotated as gut epithelial proteins, and some of them were found to be targeted by bacterial proteins. The role of some of the targeted hubs in gut epithelia and interference with humoral immunity pathways is defined in forthcoming text.

The targeted receptor highly connected hub (degree 43) Rack1 (Cflo_N_g1100) is a multifunctional protein participates in regulating several cell surface receptors and intracellular protein kinases (Choi et al., 2003). Rack1 can also bind with integrins, which further can interact with viruses (Albinsson and Kidd, 1999) and produces signals required for the organization of actins in the cytoskeleton (Buensuceso et al., 2001; Liliental and Chang, 1998). Additionally, Rack1 contributes to NF-κB mediated cell survival (Choi et al., 2003). The high connectivity of Rack1 in the IC with diverse proteins clue the binding and signal propagation capacity of Rack1 for modulating different physiological functions.

Notch signalling plays a central role in metazoan development, epithelial stem cell maintenance and controlling the gut homeostasis and host physiology (Fre et al., 2011; Wang et al., 2014). The Notch receptor (Cflo_N_g6260) was found to be up-regulated after immune

challenge. On depletion of Notch in the progenitor cells, the host fails to survive during sublethal pathogenic gut-microbe interactions. Interestingly, the differentiation of the intestinal stem cells (ISCs) via Notch and other humoral immunity pathways (Jak-Stat and JNK signalling) is also an essential event for host resistance against bacteria pathogens (Jiang et al., 2009; Lee, 2009). Therefore, it is conceivable that the hampering of Notch by bacterial proteins may activate the microbe-induced epithelial renewal program by stimulation of Notch signalling pathway.

Some of the semaphorins are referred as immune semaphorins due to their ability to enhance host immunity and manipulating pathological immune responses (Kikutani and Kumanogoh, 2003; Kumanogoh and Kikutani, 2003; Mizui et al., 2009). Immune semaphorins such as Sema4D, functions as a receptor in the immune system (Kumanogoh and Kikutani, 2003). The interactions of semaphorin Sema1A (Cflo_N_g8419) with bacterial proteins was identified. Sema1A and Sema4D both consists the similar SEMA domain and cysteine rich plexin repeat region however, the additional immunoglobulin domain of Sema4D lacks in Sema1A.

The homolog of Patj-like protein (Patj; Cflo_N_g6074) is required for the formation of tight junctions (TJ) in mammalian epithelial cells (Shin et al., 2005). The role of PDZ domain-containing protein Patj and bacterial interaction is yet not explored, however the TJ-associated Patj is a binding partner and degradation target of E6 protein of high-risk human papillomavirus (HPV) types 16 and 18 (Storrs and Silverstein, 2007).

The *D. melanogaster* scribble protein (Scrib) localizes at the septate junction through LRRs and PDZ domains, which is functionally identical to the vertebrate TJ, (Bilder and Perrimon, 2000) and can negatively regulate the cell proliferation by localizing at the basolateral membrane of epithelial cells (Nagasaka et al., 2006). Additionally, this protein is also involved in cell-cycle regulation (Nagasaka et al., 2006) and MAPK signalling (Dow et al., 2008). The analysis indicated the physical interaction of ant Scrib (Cflo_N_g8357) with Serratia type 1 secretion system membrane fusion protein HASE.

3.17.2.1 Interaction with IMD pathway members

In the Illumina sequencing only a few genes in the IMD pathway showed the differential expression on immune challenge despite the elevated production of AMPs, among which PGRP-LB and a pirk-like protein are the inhibitor of the IMD pathway. This gives an

indication that there might be other routes for activation of IMD pathway. The disruption of NF-κB signalling by the physical interactions of pathogen proteins have shown in *D. melanogaster* (Lindmark et al., 2001; Thoetkiattikul et al., 2005) and human by yeast two-hybrid assay (Dyer et al., 2010). In host-pathogen interaction analysis bacterial proteins targeting the apoptosis inhibitor 2 (Iap2; Cflo_N_g2123) and NF-κB Rel (Cflo_N_g6082) were identified. Iap2 act as a regulator of IMD signalling (Gesellchen et al., 2005; Kleino et al., 2005) and involves in positive regulation of Fadd (Guntermann and Foley, 2011) and Dredd (Meinander et al., 2012) which further activates Rel in downstream signalling cascade (Gupta et al., 2015). The activated Rel ultimately induces the expression AMP gene hymenoptaecin (Ratzka et al., 2013; Schlüns and Crozier, 2007) for elimination of bacterial infection.

3.17.2.2 Interaction with JNK pathway members

Rho GTPases (Rac1 and Rho1) the upstream components of the JNK pathway have shown the modulation of their activity after targeting by several secreted proteins of bacterial pathogens (Burridge and Wennerberg, 2004; Richard et al., 2010). Moesin (Moe; Cflo_N_g540), the upstream regulator of Rho1 (Neisch et al., 2010) was also found to be targeted by bacterial proteins. Moreover, apoptosis inhibitor 1 (Iap1; Cflo_N_g2127) and a JNK activator MAPKKK kinase Ask1 (Cflo_N_g10679) were also found in interaction with bacterial proteins. In JNK cascade the Iap1 inhibits Traf1 which is an activator of Ask1 that act as an inducer of JNK in the downstream pathway. Interestingly, similar to Rel of IMD signalling the AP-1 family TF Jun (Cflo_N_g3291) and Fos (Cflo_N_g4719) were found to be targeted by bacteria. The RNA interference (RNAi) knockdown of FOS in S2 cells of *D. melanogaster* can block the expression of AMPs attacin and drosomycin (Kallio et al., 2005). The Jun loss-of-function studies in vivo suggest the role of these AP-1 TFs in control of AMPs gene expression (Leppä and Bohmann, 1999). Together, these results of host-pathogen interactions suggest that *Serratia* could modulate the JNK pathway activity although the fate of interaction is yet to be investigated in future. In agreement with these finding, a recent study has suggested the JNK pathway are functional in aphid defence against *Serratia* and recommended to attentively explore the role of JNK pathway in insect-bacterial interactions (Renoz et al., 2015).

3.17.2.3 Interaction with Jak-Stat pathway

Jak-Stat signalling pathway is involved in control of host defence in the gut in regulating stem cell proliferation, development and expression of thioester containing proteins (TEPs). The pathway has shown the importance in *D. melanogaster* immune responses during *S. marcescens* infection (Cronin et al., 2009). However, how this pathway is activated in social insects against the bacterial pathogens is unknown (Evans et al., 2006; Gupta et al., 2015). The host-pathogen reported here indicate the direct interference of bacterial proteins with Jak-Stat cascade which probably may trigger the pathway. Hopscotch (Hop; Cflo_N_g4220) is one of the core components of pathway found to be targeted by proline dipeptidase (protease PEPQ). Stam1 (Cflo_N_g11263), the stimulator of Hop interacts with *Serratia* RCSD, a phosphotransfer intermediate in two-component regulatory system RcsBC and *Serratia* ribonucleoside diphosphate reductase NRDB. The pathway inhibitor Ken (Cflo_N_g7785) was found to be in interaction with LPTA, a lipopolysaccharide export system protein. Furthermore, the interaction between Stat5B (Cflo_N_g13035) the homolog of *D. melanogaster* Stat92E and multiple bacterial proteins including protease HTRA, hydrolase KIPA, autotransporter assembly complex protein TAMB, and lipopolysaccharide export system protein LPTC was also identified. Moreover, the multiligand recognition protein Tep1 (Cflo_N_g7345) which is a product of Jak-Stat activation, was found to be in interaction with an extracellular leucine-rich repeat protein SMDB11_2039 and a cytoplasmic membrane protein MTLA, a component of mannitol-specific phosphotransferase system. SMDB11_2039 shows 52 % sequence identity with *Yersinia nurmii* YOPM like virulence protein encoded by small leucine-rich proteoglycan (SLRP). Notably, such proteoglycans can act as a damage-associated molecular patterns (DAMPs) in innate immunity (Moreth et al., 2012; Schaefer and Iozzo, 2008). TEPs have been directly implicated as opsonins mediating the insect innate immune response by promoting phagocytosis of pathogenic bacteria (Bou Aoun et al., 2010; Levashina et al., 2001). TEP1 also contains the canonical thioester motif (CGEQ) which is important for the formation of the covalent bond to the bacterial surface (Bou Aoun et al., 2010; Janssen et al., 2005).

3.17.2.4 Interaction with Toll Pathway

Gram-positive bacteria, fungi and other virulence factors often activate the Toll signalling pathway. Interestingly, the host-bacteria PPI here showed that few proteins in the Toll pathway also interact with bacteria, including Fadd (Cflo_N_g10862) and NF-κB protein

Cactus (Cflo_N_g14413) although the role of the Toll pathway in the defence against Gram-negative bacteria is not well documented.

3.17.3 Interaction with cytoskeleton proteins

Bacterial pathogens have central tendency to subvert the host cell cytoskeleton to facilitate their survival, replication, and dissemination (Bhavsar et al., 2007; Haglund and Welch, 2011). The studies have shown that the components of the cytoskeleton (actin, microtubules, intermediate filaments and septins), have not only for their roles in cell division, shape, and movement but also have important functions in innate immunity and cellular self-defence (Haglund and Welch, 2011; Mostowy and Shenoy, 2015). In the host-pathogen PPI, most the cytoskeletal proteins and their binders were found to be in interaction with bacterial proteins. In the actin subnetwork, the actin constituents (A4, Arp3, Cadf, Cap1); annotated actin binders (Hrp65, Flii, Vinc), proteins involved in actin cytoskeleton reorganization (Rac1, Flnc, Pkn) and a regulator of actin organization (Rho1) were targeted by *Serratia*. Similarly, in the microtubule subnetwork, the microtubule constituents (αTub84B, TUBB4A, Dctn2, Clip1, Ndel1A); annotated microtubule binders (Mapre1, Klc1, Khk, Cnn, Hk, Ndk), protein involved in microtubule reorganization (Bicd) and proteins involved in microtubule polarization (Rac1, Prkar1B) were targeted. Additionally, some structural constituent of microtubules (Tuba1A, Tuba1B, Tuba1C, Tuba3C), actin-associated heterogeneous nuclear ribonucleoprotein (Hnrnph2) and a regulator of actin organization (Gek) were not present in IC but they were found to be in interaction with bacterial proteins. Additionally, some other proteins with cytoskeletal localization such as Epb41L1, Ttn, and Myh14 were targeted by bacteria. The disorganization of the cytoskeleton has also been uncovered by inoculation of *S. marcescens* cytotoxins in cultured Chinese hamster ovary (CHO) cells (Carbonell et al., 2004).

3.17.4 Interactions with extracellular matrix (ECM) proteins

The multifunctional extracellular matrix (ECM) of insect also provides signalling cues that regulates cell behaviour and function in tissue genesis and homeostasis (Daley and Yamada, 2013; Lukashev and Werb, 1998). Here, the interactions of bacterial proteins with host laminin, collagen IV, integrin beta Itgb, integrin-linked protein kinase-like protein (Ilk), Adam10 and Adamts9 were identified.

3.17.5 Interactions with haemolymph proteins

Several phagocytosis mediator proteins involve in interaction with bacterial proteins which include the haemocyte receptors, immunoglobulin superfamily proteins, cytoskeleton proteins, complement binding protein TEP and lysosomal enzymes. The scavenger receptor fasciclin-2 (Feel2) was found to be targeted by bacterial flagellar export component FLHB and *Y. pestis* multi-pass transmembrane protein Y3220 like permease. Feel2 also interacts with 36 other proteins in IC, which suggest that the topologically important Feel2 can interfere with several pathways including the defence mechanisms against bacterial infection. The cytoskeletal associated hubs in IC as actin protein (A4) with degree 19, actin-related protein 2/3 complex subunit 4 (Arpc4) with degree 11 and F-actin-capping protein subunit alpha (Cap1) also interacts with bacterial proteins and annotated to have involvement in phagocytosis as identified by GO terms. The bacterial targeted high degree Ras-related protein (Rac1) was annotated as receptor that participates in immune signalling, phagocytosis, and encapsulation. The high degree hub, microtubule-associated protein RP/EB family member 1-like protein (Mapre1) that interacts with 40 other proteins was found to be hampered by bacterial proteins. Mapre1 belongs to GO functional category GO:0035011 i.e., melanotic encapsulation of foreign target. Interaction of complement receptor sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein (Svep1) with a leucine-rich repeat contacting bacterial protein (SMDB11_2039) were also found. The Ras-like GTP-binding protein (Rho1) mediates 15 interactions in IC, was targeted by bacterial multi-pass membrane protein (Cora) and was annotated to be involved in melanisation defence response (GO:0035006). Interestingly, the trypsin-1-like serine protease (Try1) interacts with bacterial chitinase (CHIC). try1 gene was found to be down-regulated during bacterial infection in *C. floridanus* as shown by Illumina sequencing analysis, and participates in melanisation responses against pathogens (Chu et al., 2015). The putative interaction between mitochondrial serine protease (Htra2) of ant and three bacterial proteins were identified. Htra2 contributes to the efficient induction of apoptosis (Martins et al., 2002). It was observed, besides TEP1, three other microbial recognition receptors namely galectin-8 (Lgals8), multiple epidermal growth factors 8 like protein (Megf8) and a croquemort (crq) family class B scavenger receptor (Scarb1) interacts with bacterial proteins. The *C. floridanus* Megf8 which is the ortholog of *N. vitripennis* microbial recognition protein Nasvi2EG006946 (Sackton et al., 2013) interacts with a bacterial transporter protein PROW. The scavenger receptors are multi-ligand receptors that have emerged as important PRRs because of their ability to bind a wide range of microbes including bacteria (Stuart et al., 2005) and viruses

(Zeisel et al., 2007). The cytosolic galectin Lgals8 act as a danger receptor that induces phagocytic clearance of endocytosed or phagocytosed pathogens and triggers the antibacterial autophagy-initiating machinery (Thurston et al., 2012). The PPI of Lgals8 with bacterial chemotaxis regulator CHEY and a secretory protein SYD which function as immune-related protein in bacterial toxin systems were identified. Additionally, orthologues of mammalian Scarb1 also exist in *C. floridanus* which interacts with three bacterial carboxypeptidases (Dacd, Smdb11_Rs04510, Smdb11_Rs09345) and a divalent ion tolerance protein CUTA in the host-pathogen interactome.

3.17.6 Interactions with proteins functionally associated with cuticle

Earlier studies have suggested that the cuticle can also participate actively in immune defence in silkworm (Brey et al., 1993). Seven plausible interactions between the cuticle associated proteins and bacteria were identified. All the proteins defined here mediates in some cutical associated functions such as development or chitin based cuticle or regulation of adult chitin-containing cuticle pigmentation or molting cycle. Node cAMP-dependent protein kinase type I regulatory subunit (Prkar1B) which interacts with 13 other proteins in IC found to interact with bacterial proteins. The ortholog of jagged family protein (SER) of *Camponotus* is also targeted by *Serratia* which is known to involve in hemopoiesis (Lebestky et al., 2003). Another bacterial target LDL receptor-related protein 1 (Lrp1) is a large multidomain containing multifunctional scavenger receptor (Herz and Strickland, 2001) which has also been shown to act downstream of complement-like pathway activation and mediating phagocytosis by binding with TEP1-opsonized bacteria (Moita et al., 2005). Moreover, the 26S proteasome regulatory subunit S3 (Dox-A2) was also identified as bacterial target, which is indeed not a cuticular protein but involves in melanin formation and sclerotization of the cuticle (Garvey and Malcolm, 2000).

3.18 Targets of virulence associated *S. marcescens* proteins in *C. floridanus* interactome

Many extracellular enzymes released by *S. marcescens*, for instance, nucleases (Hejazi and Falkiner, 1997), hemolysins, proteases (Matsumoto et al., 1984) and cytotoxins affects the host cell lines (Kuehn and Kesty, 2005). Together with the type-VI secretion system effectors, serralysins (Zhang et al., 2015) and hemolysins (Hertle et al., 1999) are the key virulence factors of *Serratia* to survive intracellularly and invade the host defence. The virulence factors such as fimbriae helps bacteria to for adhere with host cells. Using Dikshitra

algorithm a subnetwork of *C. floridanus* proteins targeted by such was extracted factors (Fig. 17).

Type VI secretion system VgrG-like proteins targets many host proteins including the high degree high density lipoprotein-binding protein vigilin (HDLBP) and cytoskeletol protein kinesin light chain 1 (Klc1). The host NF-κB Rel was found to be targeted by bacterial type VI secretion system-associated ImpC like protein and other six fimbrial proteins. Interestingly, the ability of fimbriae to indirectly activate NF-κB has been shown in human monocytic cells (Hajishengallis and Genco, 2004) although here the analysis suggested direct interference of fimbrial proteins with REL. Hemolysin ShlA targets the hub DEAD (Asp-Glu-Ala-Asp) box polypeptide 39 protein Ddx39 which is a growth-associated RNA helicase that mainly function in RNA metabolism and ubiquitin-proteasome degradation pathway (Sugiura et al., 2007). All the serralysins were found to be in interactions with three ant proteins Tm9sf2, Stk11 and Acly. Tm9sf2, the paralogue of Tm9sf4 belong to transmembrane 9 superfamily and together with Tm9sf4 synergistically promote the Gram-negative bacterial phagocytosis in *Drosophila* S2 cells, actin cytoskeleton network reorganization, expression of cell surface proteins, membrane receptor trafficking and regulation of signalling activity (Bergeret et al., 2008). Both Tm9sF2 and Tm9sf4 co-localize and interact with PGRP-LC receptor and exert a regulatory function at the signalling activity of PGRP-LC (Perrin et al., 2014). The serralysin targeted Tm9sf2 is implicated in preventing inappropriate signalling from the unstimulated PGRP-LC receptor while Tm9sf4 participates in PGRP-LC localization at the plasma membrane (Perrin et al., 2014).

Stk11, another preferred target of serralysins, is a pleiotropic serine/threonine kinase that displays numerous functions in epithelial cells, including its requirement for efficient dissemination of Gram-negative bacterium Shigella flexneri in epithelial cells (Dragoi and Agaisse 2014). Furthermore, the interactions of serralysins with ATP citrate lyase (Acly) was identified which often involves in lipogenesis, tumorigenesis, inflammatory metabolism and cellular senescence in human cells (Infantino, Iacobazzi et al. 2013, Lee, Jang et al. 2015).

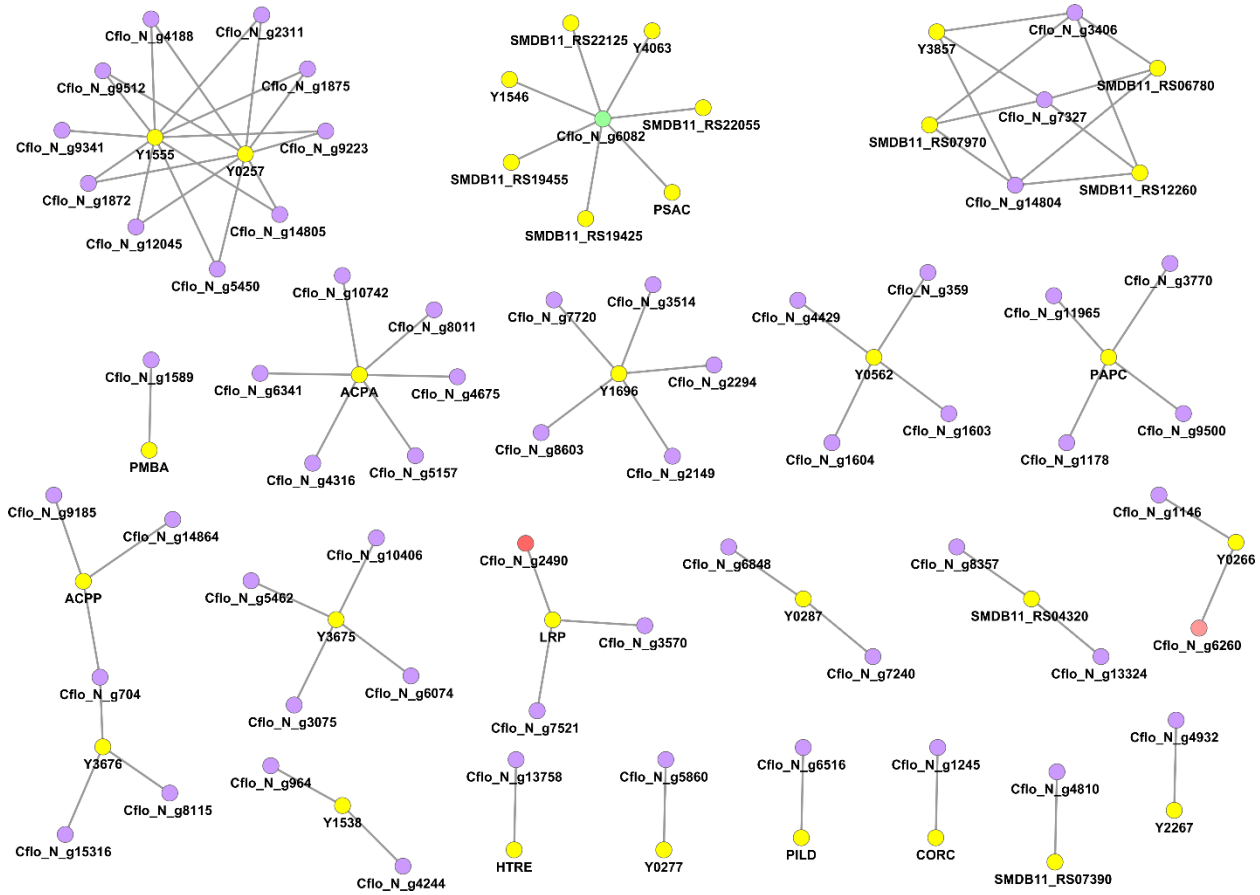
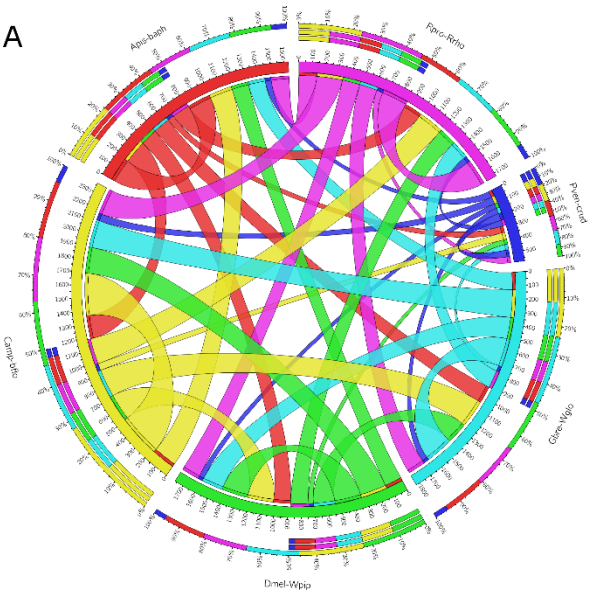


Figure 17. Host-pathogen interactions between *C. floridanus* proteins and *Serratia* virulence related proteins. Node designation: Yellow, *Serratia* virulence related proteins; red gradients, down-regulated genes; green gradients, up-regulated genes; violet, not differentially expressed. All nodes are denoted by re-annotation identifiers (see Appendix VI for annotation and expression value). Connectivity among nodes is based on undirected interactions (—).

3.19 Interaction of *C. floridanus* with endosymbiont *Blochmannia floridanus*

The analysis of host-endosymbiont PPIs resulted in 327 final interactions in *C. floridanus* and *B. floridanus*. The conservation score for an interaction is defined as a total number of host-endosymbiont pairs in which interactions are found. Using the *C. floridanus - B. floridanus interactions* as a reference, only 10 % of interactions were found to be highly conserved in all the six insect-endosymbionts pairs. *C. floridanus - B. floridanus* shares the highest number of interactions with *G. brevipalpis - W. glossinidia* and lowest number of interactions with while that of the lowest with *P. venusta - C. ruddii* pair (Fig. 18). This could be attributed by the fact that the endosymbiont *B. floridanus* is phylogenetically closer with *W. glossinidia* (Belda et al., 2005; Gil et al., 2003) and the *C. ruddii* has smallest genome among the analysed species (Nakabachi et al., 2006).

A



B

	Camp-bflo	Apis-baph	Pven-crud	Rpro-Rrho	Gbre-Wglo	Dmel-Wpip
Camp-bflo	327					
Apis-baph	182	182				
Pven-crud	55	49	55			
Rpro-Rrho	200	139	47	200		
Gbre-Wglo	215	133	44	139	215	
Dmel-Wpip	204	132	38	138	166	204

Figure 18. Interologs in six host-endosymbiont protein-protein interaction networks. Interactions between *C. floridanus* and *B. floridanus* was used as reference for the analysis. (A) Circos plot representation of the interolog conservancy. Thickness of coloured the strips displays the graphical visualization of the number of shared interologs. (B) Quantitative details of the interolog conservancy. Abbreviation: 'Apis' - *Acyrthosiphon pisum*, 'Baph' - *Buchnera aphidicola*, 'Pven' - *Pachypsylla venusta*, 'Crud' - *Carsonella ruddii*, 'Rpro' - *Rhodnius prolixus*, 'Rrho' - *Rhodococcus rhodnii*, 'Gbre' - *Glossina brevipalpis*, 'Wglo' - *Wigglesworthia glossinidia*, 'Dmel' - *D. melanogaster*, 'Wpip' - *Wolbachia pipientis wMel*.

Several *B. floridanus* proteins were found be in interaction with IMD, Jak-Stat, and JNK pathways. IMD pathway is the major regulator of Gram-negative pathogenic bacteria and may also contribute to control and tolerance of the endosymbiont (Douglas et al., 2011; Ratzka et al., 2013). The analysis of immune specific interactome revealed the NF-κB factor mammalian NF-κB2 homolog in *C. floridanus* Rel as top-ranked hub in *C. floridanus* - *B. floridanus* interspecies network. The interactions between Rel and *B. floridanus* proteins including eight membrane proteins, two proteins with non-classical secretion signals, one virulence factor trpD (Movahedzadeh et al., 2008; Smith et al., 2001) and one non-classical virulence related protein purB (Ge et al., 2008; Okazaki et al., 2007) were also predicted. The expression of Rel in *C. floridanus* increases from larvae to workers in the body, while it was expressed significantly lower in the midgut of pupae than either worker or larvae (Ratzka et al., 2013). The Rel targeting by *B. floridanus* likely impact the regulation of immune response. However, these interactions were not highly conservancy in the other host-endosymbionts pairs, which strengthen the speculation that the traditional view of insect immunity could not be broadly applicable (Gerardo et al., 2010), at least for endosymbiont bearing insects. In JNK pathway Rho1, Rac1, Kay, and MAPKKK15 were found to be in interaction with *Blochmannia* proteins. Moreover, in Jak-Stat signalling Ken, Stam1 and Stat

were found as targets of *B. floridanus* proteins. Interestingly, the control of endosymbiont by Jak-Stat pathway is yet not explored and opens the new avenues for insect biologists. The comparison of virulence associated *S. marcescens* proteins – *C. floridanus* host pathogen interactome with host – endosymbiont network revealed the presence of interolog PPIs between two virulence associated *Serratia* protein and ant proteins. These *Serratia* proteins includes an acyl carrier protein (AcpP) and a metalloprotease (PmbA-encoded protein). The hemolytic activity of AcpP towards host is shown in *Brachyspira hyodysenteriae* (Hsu et al., 2001) while PmbA encoded protein MccB17 is classically documented as a bacterial toxin capable to inhibit the DNA replication (Vizan et al., 1991). Altogether, the analysis of host - endosymbiont network suggests that similar to pathogen, *B. floridanus* could also interfere with the *C. floridanus* immune system.

3.20 Haemolymph immunity in larvae and adults

Pools of haemolymph proteins from *C. floridanus* larval and adults were collected separately from control and immune challenged ants. Both the larvae and adults haemolymph proteins were compared with corresponding untreated samples to identify the differentially expressed proteins (DEPs). Among the 1291 proteins identified by MS, 192 proteins were identified as differentially expressed on immune challenge in larvae, adult or in both.

Table 7. GO terms significantly enriched in the set of proteins that are differentially expressed in *C. floridanus* larvae haemolymph upon immune challenge. Category F represents molecular function, category P represents biological process and category C represents cellular compartments.

GO-ID	Term description	Category	P-value	Significance
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	F	6.63E-05	***
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	F	4.80E-04	***
GO:0016798	hydrolase activity, acting on glycosyl bonds	F	9.52E-04	***
GO:0005198	structural molecule activity	F	4.31E-03	**
GO:0005509	calcium ion binding	F	4.97E-03	**
GO:0030246	carbohydrate binding	F	4.25E-03	**
GO:0016787	hydrolase activity	F	1.91E-02	*
GO:0008061	chitin binding	F	1.46E-02	*
GO:0005200	structural constituent of cytoskeleton	F	2.94E-02	*
GO:0071855	neuropeptide receptor binding	F	1.04E-02	*
GO:0016657	oxidoreductase activity, acting on NAD(P)H, nitrogenous group as acceptor	F	1.04E-02	*

GO:0048408	epidermal growth factor binding	F	1.04E-02	*
GO:0060308	GTP cyclohydrolase I regulator activity	F	1.04E-02	*
GO:0019010	farnesoic acid O-methyltransferase activity	F	2.07E-02	*
GO:0003933	GTP cyclohydrolase activity	F	2.07E-02	*
GO:0035049	juvenile hormone acid methyltransferase activity	F	2.07E-02	*
GO:0009982	pseudouridine synthase activity	F	3.09E-02	*
GO:0008171	O-methyltransferase activity	F	3.09E-02	*
GO:0004367	glycerol-3-phosphate dehydrogenase [NAD+] activity	F	4.10E-02	*
GO:0004360	glutamine-fructose-6-phosphate transaminase (isomerizing) activity	F	4.10E-02	*
GO:0042562	hormone binding	F	4.10E-02	*
GO:0005975	carbohydrate metabolic process	P	6.25E-03	**
GO:0006022	aminoglycan metabolic process	P	2.60E-03	**
GO:0006026	aminoglycan catabolic process	P	1.55E-03	**
GO:0046907	intracellular transport	P	1.71E-02	*
GO:0015031	protein transport	P	4.63E-02	*
GO:1901071	glucosamine-containing compound metabolic process	P	2.33E-02	*
GO:0031032	actomyosin structure organization	P	2.33E-02	*
GO:0006040	amino sugar metabolic process	P	2.53E-02	*
GO:0007498	mesoderm development	P	4.07E-02	*
GO:0043095	regulation of GTP cyclohydrolase I activity	P	1.04E-02	*
GO:0030433	ER-associated ubiquitin-dependent protein catabolic process	P	1.04E-02	*
GO:0051703	intraspecies interaction between organisms	P	1.04E-02	*
GO:0035185	preblastoderm mitotic cell cycle	P	1.04E-02	*
GO:0035176	social behaviour	P	1.04E-02	*
GO:0010632	regulation of epithelial cell migration	P	1.04E-02	*
GO:0019464	glycine decarboxylation via glycine cleavage system	P	1.04E-02	*
GO:0018990	ecdysis, chitin-based cuticle	P	1.04E-02	*
GO:0030970	retrograde protein transport, ER to cytosol	P	1.04E-02	*
GO:0006728	pteridine biosynthetic process	P	2.07E-02	*
GO:0046654	tetrahydrofolate biosynthetic process	P	2.07E-02	*
GO:0019889	pteridine metabolic process	P	2.07E-02	*
GO:0044130	negative regulation of growth of symbiont in host	P	2.07E-02	*
GO:0016114	terpenoid biosynthetic process	P	3.09E-02	*
GO:0006013	mannose metabolic process	P	3.09E-02	*
GO:0032509	endosome transport via multivesicular body sorting pathway	P	4.10E-02	*
GO:0045746	negative regulation of Notch signalling pathway	P	4.10E-02	*
GO:0046168	glycerol-3-phosphate catabolic process	P	4.10E-02	*
GO:0005576	extracellular region	C	3.61E-04	***
GO:0044444	cytoplasmic part	C	9.02E-03	**
GO:0044421	extracellular region part	C	8.56E-03	**
GO:0005737	cytoplasm	C	2.33E-02	*
GO:1990204	oxidoreductase complex	C	1.31E-02	*
GO:0030117	membrane coat	C	2.33E-02	*
GO:0048475	coated membrane	C	2.33E-02	*
GO:0031395	bursicon neuropeptide hormone complex	C	1.04E-02	*
GO:0030864	cortical actin cytoskeleton	C	3.09E-02	*

GO:0044420	extracellular matrix part	C	4.10E-02	*
GO:0036452	ESCRT complex	C	4.10E-02	*

In total, 11 proteins showed the differential expression both in larvae and adults among which Niemann-Pick disease type C2 (NPC2) is recognised as microbial receptor (Shi et al., 2012). The over-expression of NPC2 can stimulate the expression of AMPs, probably via IMD pathway (Shi et al., 2012). The over-expression of NPC2 is identified in *A. mellifera* (Richard et al., 2012) and *D. melanogaster* (Shi et al., 2012) after bacterial challenge. GO annotation of larval DEPs resulted in over-representation of 59 GO terms (Table 7) which include the enrichment of proteins involved in hydrolase activities (molecular function), carbohydrate and aminoglycan metabolic process (biological process) and extracellular region (cellular compartment) as top GO categories.

Moreover, GO annotation of adult DEPs resulted in over-representation of 84 GO terms (Table 8) which include the enrichment of proteins involved in oxidoreductase activity (molecular function), single-organism metabolic process (biological process) and oxidoreductase complex (cellular compartment) as top GO categories.

In all the DEPs, over-representation of only few GO terms occurred in both the larvae and adults. In molecular function category, proteins involved in calcium ion binding (GO:0005509) and glycerol-3-phosphate dehydrogenase [NAD+] activity (GO:0004367) were over-represented in both the larvae and adults. Furthermore, in biological process category, glycerol-3-phosphate catabolic process (GO:0046168) was over-represented in both the larvae and adults, and in cellular compartment category notably oxidoreductase complex (GO:1990204) were over-represented in both the larvae and adults. However, in adults several term related to oxidoreductase activity were enriched (Table 8) which indicates the role of this category is specifically attributed to adults.

Table 8. GO terms significantly enriched in the set of proteins that are differentially expressed in *C. floridanus* adults haemolymph upon immune challenge. Category F represents molecular function, category P represents biological process and category C represents cellular compartments.

GO-ID	Term description	Category	P-value	Significance
GO:0016491	oxidoreductase activity	F	1.41E-04	***
GO:0009055	electron carrier activity	F	2.97E-06	***
GO:0016209	antioxidant activity	F	1.51E-04	***

GO:0046872	metal ion binding	F	6.36E-03	**
GO:0043169	cation binding	F	6.97E-03	**
GO:0005509	calcium ion binding	F	1.97E-03	**
GO:0050660	flavin adenine dinucleotide binding	F	1.75E-03	**
GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	F	1.42E-03	**
GO:0004601	peroxidase activity	F	1.42E-03	**
GO:0016667	oxidoreductase activity, acting on a sulfur group of donors	F	3.64E-03	**
GO:0008483	transaminase activity	F	9.76E-03	**
GO:0016769	transferase activity, transferring nitrogenous groups	F	9.76E-03	**
GO:0004352	glutamate dehydrogenase (NAD+) activity	F	8.43E-03	**
GO:0004333	fumarate hydratase activity	F	8.43E-03	**
GO:0045155	electron transporter, transferring electrons from CoQH2-cytochrome c reductase complex and cytochrome c oxidase complex activity	F	8.43E-03	**
GO:0004148	dihydrolipoyl dehydrogenase activity	F	8.43E-03	**
GO:0004020	adenylylsulfate kinase activity	F	8.43E-03	**
GO:0004001	adenosine kinase activity	F	8.43E-03	**
GO:0008177	succinate dehydrogenase (ubiquinone) activity	F	8.43E-03	**
GO:0004781	sulfate adenylyltransferase (ATP) activity	F	8.43E-03	**
GO:0048037	cofactor binding	F	1.13E-02	*
GO:0020037	heme binding	F	1.20E-02	*
GO:0046906	tetrapyrrole binding	F	1.57E-02	*
GO:0051540	metal cluster binding	F	3.11E-02	*
GO:0051536	iron-sulfur cluster binding	F	3.11E-02	*
GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	F	3.65E-02	*
GO:0016721	oxidoreductase activity, acting on superoxide radicals as acceptor	F	1.68E-02	*
GO:0004736	pyruvate carboxylase activity	F	1.68E-02	*
GO:0003878	ATP citrate synthase activity	F	2.51E-02	*
GO:0016885	ligase activity, forming carbon-carbon bonds	F	2.51E-02	*
GO:0004775	succinate-CoA ligase (ADP-forming) activity	F	2.51E-02	*
GO:0004367	glycerol-3-phosphate dehydrogenase [NAD+] activity	F	3.33E-02	*
GO:0043022	ribosome binding	F	3.33E-02	*
GO:0004075	biotin carboxylase activity	F	3.33E-02	*
GO:0016638	oxidoreductase activity, acting on the CH-NH2 group of donors	F	4.15E-02	*
GO:0046912	transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	F	4.15E-02	*
GO:0043021	ribonucleoprotein complex binding	F	4.96E-02	*
GO:0044710	single-organism metabolic process	P	1.87E-04	***
GO:0055114	oxidation-reduction process	P	5.44E-07	***
GO:0006091	generation of precursor metabolites and energy	P	6.62E-06	***
GO:0009060	aerobic respiration	P	1.55E-05	***
GO:0006106	fumarate metabolic process	P	6.95E-05	***
GO:0072593	reactive oxygen species metabolic process	P	1.88E-03	**
GO:0043648	dicarboxylic acid metabolic process	P	5.11E-03	**
GO:0045454	cell redox homeostasis	P	7.74E-03	**
GO:0006734	NADH metabolic process	P	8.43E-03	**

GO:0000103	sulfate assimilation	P	8.43E-03	**
GO:0043174	nucleoside salvage	P	8.43E-03	**
GO:0006452	translational frameshifting	P	8.43E-03	**
GO:0008364	pupal chitin-based cuticle development	P	8.43E-03	**
GO:0006127	glycerophosphate shuttle	P	8.43E-03	**
GO:0006116	NADH oxidation	P	8.43E-03	**
GO:0045727	positive regulation of translation	P	8.43E-03	**
GO:0007320	insemination	P	8.43E-03	**
GO:0006082	organic acid metabolic process	P	3.50E-02	*
GO:0016051	carbohydrate biosynthetic process	P	1.09E-02	*
GO:0044282	small molecule catabolic process	P	2.44E-02	*
GO:0046033	AMP metabolic process	P	1.68E-02	*
GO:0042743	hydrogen peroxide metabolic process	P	1.68E-02	*
GO:0034614	cellular response to reactive oxygen species	P	1.68E-02	*
GO:0006121	mitochondrial electron transport, succinate to ubiquinone	P	1.68E-02	*
GO:0008612	peptidyl-lysine modification to hypusine	P	2.51E-02	*
GO:0046516	hypusine metabolic process	P	2.51E-02	*
GO:0006937	regulation of muscle contraction	P	2.51E-02	*
GO:0046168	glycerol-3-phosphate catabolic process	P	3.33E-02	*
GO:0006662	glycerol ether metabolic process	P	4.96E-02	*
GO:0043094	cellular metabolic compound salvage	P	4.96E-02	*
GO:1990204	oxidoreductase complex	C	3.50E-04	***
GO:0044444	cytoplasmic part	C	3.44E-03	**
GO:0005739	mitochondrion	C	7.84E-03	**
GO:0030017	sarcomere	C	1.58E-03	**
GO:0043292	contractile fiber	C	1.75E-03	**
GO:0005759	mitochondrial matrix	C	3.29E-03	**
GO:0070469	respiratory chain	C	3.64E-03	**
GO:0045273	respiratory chain complex II	C	8.43E-03	**
GO:0045251	electron transfer flavoprotein complex	C	8.43E-03	**
GO:0005749	mitochondrial respiratory chain complex II	C	8.43E-03	**
GO:0017133	mitochondrial electron transfer flavoprotein complex	C	8.43E-03	**
GO:0005737	cytoplasm	C	2.00E-02	*
GO:0005856	cytoskeleton	C	4.07E-02	*
GO:0015629	actin cytoskeleton	C	1.84E-02	*
GO:0005811	lipid particle	C	1.71E-02	*
GO:0045239	tricarboxylic acid cycle enzyme complex	C	1.68E-02	*
GO:0033180	proton-transporting V-type ATPase, V1 domain	C	4.15E-02	*

To identify the major regulators of *C. floridanus* immunity the MS expression data were correlated with mRNA expression data. In total, 30 significantly expressed proteins were identified from the haemolymph to have significant differential expression of their corresponding mRNA. Interestingly, only 4 proteins from adults showed the significant changes in their corresponding mRNA while for larvae expression of 28 proteins had changes

in their mRNA expression on immune challenging *C. floridanus* with bacteria. The expression level in these 28 larvae proteins and mRNA showed moderate positive correlation (R^2=0.6). Among these, IMD regulator PGRP-LB and AMP defensin-2 showed the upregulation in MS and Illumina sequencing data (Table 9).

Table 9. Expression intensity of *C. floridanus* differentially expressed transcripts and their corresponding protein label-free quantification (LFQ) intensity in haemolymph of larvae.

Accession Number	Functional annotation	Log2FC DESeq	Log2LFQ.intensity MS
Cflo_N_g8312t1	Defensin-2	4.59	3.32
Cflo_N_g5252t1	Protein G12	2.77	-2.57
Cflo_N_g103t1	PGRP-LB	2.70	1.74
Cflo_N_g8442t1	Serine protease persephone	2.50	2.03
Cflo_N_g10198t1	Translation initiation factor IF-2	2.41	2.42
Cflo_N_g4257t1	Leukocyte elastase inhibitor	2.16	2.97
Cflo_N_g4958t1	DnaJ-class molecular chaperone	1.78	4.03
Cflo_N_g8989t1	Dynein beta chain, ciliary	1.59	2.01
Cflo_N_g12187t1	Silk fibroin	1.24	3.43
Cflo_N_g13757t1	Ecdysteroid kinase	-1.31	-2.72
Cflo_N_g793t1	Calexcitin-2	-1.32	-1.70
Cflo_N_g6298t1	Zinc metalloprotease zmpB	-1.69	-3.30
Cflo_N_g7296t1	Chymotrypsin-1	-1.71	-1.75
Cflo_N_g1853t1	mucin-6-like	-1.71	2.46
Cflo_N_g5481t1	Beta-galactosidase	-1.88	-2.79
Cflo_N_g4555t1	Peritrophic membrane protein 1	-1.88	-3.86
Cflo_N_g8803t1	Zinc carboxypeptidase A 1	-1.89	-1.80
Cflo_N_g4442t1	Trypsin-1	-1.92	-4.51
Cflo_N_g2490t1	Alkaline phosphatase	-2.04	-2.46
Cflo_N_g8234t1	Maltase 1	-2.41	-3.86
Cflo_N_g1453t1	Alpha-glucosidase	-2.41	-2.75
Cflo_N_g5236t1	Protein G12	-2.48	-3.18
Cflo_N_g11635t1	Guanine deaminase	-2.56	-2.08
Cflo_N_g7901t1	Hypothetical protein	-2.70	-5.14
Cflo_N_g13662t1	Haemolymph juvenile hormone binding protein (JHBP)	-3.09	-3.80
Cflo_N_g14742t1	Guanine deaminase	-3.32	-2.26
Cflo_N_g2544t1	Trypsin-1	-3.60	-2.17
Cflo_N_g10280t1	Pancreatic triacylglycerol lipase	-5.57	2.42

3.21 Functional modules in haemolymph subnetworks

Subnetworks from IISC were created based on shortest pathway connecting the DEPs in haemolymph and further analysed for identification of active functional modules during infection. Comparatively to adults, larger active network was identified in larvae (Fig. 19).

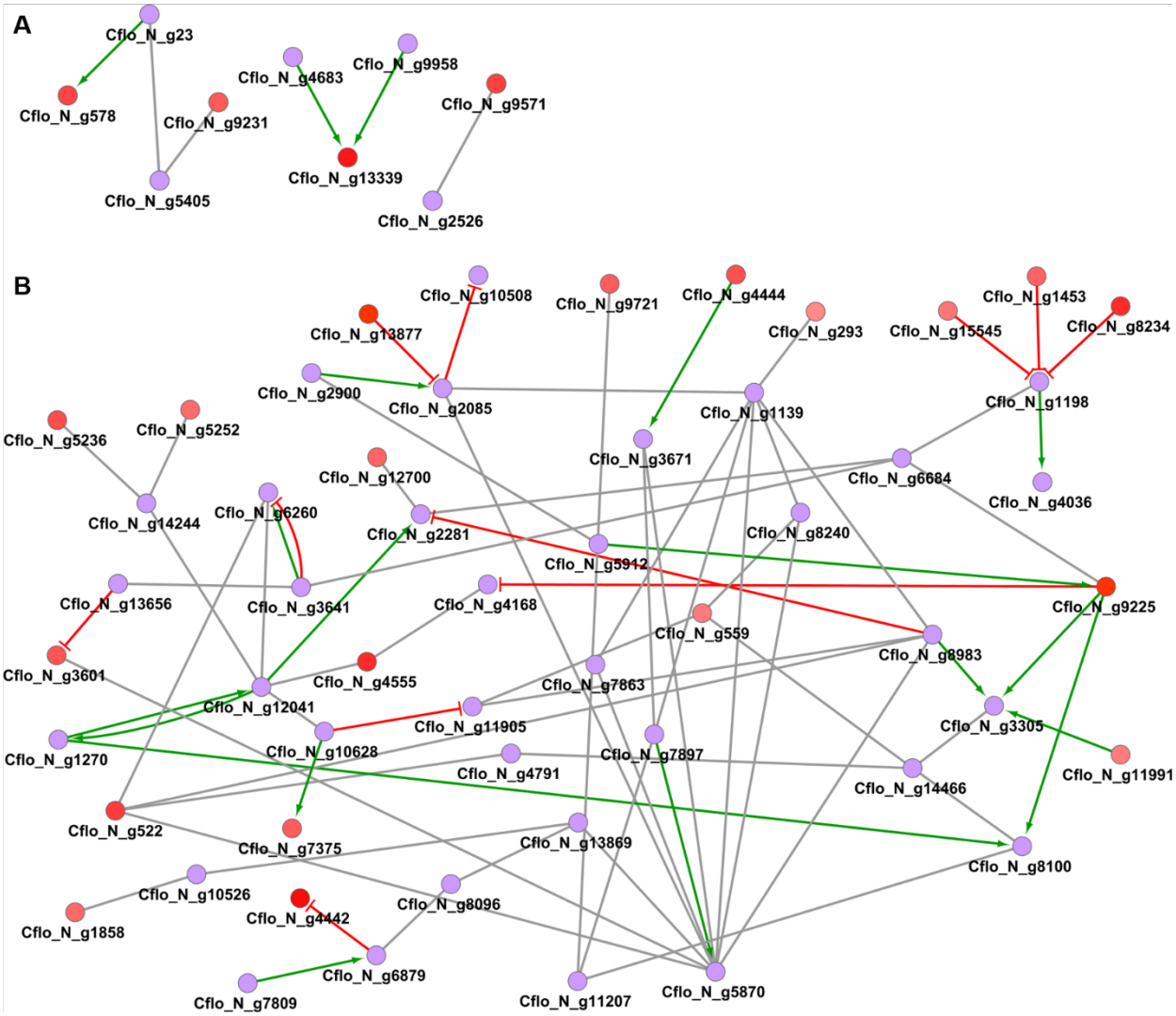


Figure 19. Merged active modules in *C. floridanus* haemolymph subnetwork. Node designation: red gradients, down-regulated proteins; green gradients, up-regulated proteins; violet, not differentially expressed. All nodes are denoted by re-annotation identifiers (see Appendix VI for annotation and expression value). Connectivity among nodes is based on undirected interactions (—). (A) Three unconnected active modules identified in adults. (B) Subnetwork generated by merging active modules in larvae.

Calcium ion dependent adhesion molecule cadherin (Cflo_N_g9231) and catenin (Cflo_N_g5405) was found in adult haemolymph active modules. Cadherin has been investigated to form a dynamic complex with catenins and regulates several intracellular

signalling pathways, including NF-κB signalling, Rho GTPase, and PI3K/Akt (Van den Bossche et al., 2012).

Another active module identified in adult haemolymph protein represents the activation of adenosine kinase 2 (Cflo_N_g13339; Ak2) by two cystathionine gamma-lyases (Cflo_N_g4683 and Cflo_N_g9958; Gcl). Ak2 and Gcl participates in methionine, transsulfuration pathway which ultimately leads to the glutathione biosynthesis. Glutathione contributes in preventing damage to cellular components caused by ROS such as free radicals, peroxides (Pompella et al., 2003). The overrepresentation of oxidoreductase activity was also identified in adults by GO enrichments. Thus, it seems this active module could be implicated in redox homeostasis during the immune response of adults after bacterial challenge.

Moreover, the smallest active module identified in adult haemolymph represent the interaction between the down-regulated extracellular matrix protein teneurin (Cflo_N_g9571) and integrin αPS2 (Cflo_N_g2526). Teneurins are signalling molecules that may function both at the cell surface as type II transmembrane receptors and, after the release of the intracellular domain, as transcriptional regulator (Tucker and Chiquet-Ehrismann, 2006). Besides regulating morphogenesis and outgrowth, the interaction of teneurin and αPS2 can promote signal transduction and intercellular adhesions (Mosca, 2015).

Integrating the active modules identified from larvae haemolymph DEPs resulted in comparatively big subnetwork (Fig. 19). Interestingly, no any up-regulated protein identified in the subnetwork. In total, 19 down-regulated proteins were connected in the merged subnetwork of active modules.

In comparison with the mRNA expression data from DESeq analysis no any overlapping active modules were identified for adults while for larvae three such modules were identified (Fig. 20). In the conserved active modules with 7 nodes, downregulation of trypsin-1 (Cflo_N_g2544) was identified. In the smallest active module identified by jActiveModules, interaction of haemolymph juvenile hormone binding protein (Cflo_N_g13662; HJHP) and lysozyme (Cflo_N_g1023) was identified. HJHP was found to be down-regulated in both the Illumina and larvae MS data. Conversely, its interacting partner lysozyme was found to be up-regulated but only in adult haemolymph MS data. Interestingly, one important active module was identified that involves in immunity of *C. floridanus*. In this module the downregulation of three cuticular proteins ultimately affect the expression level of Lysozyme c-1 (Cflo_N_g4036).

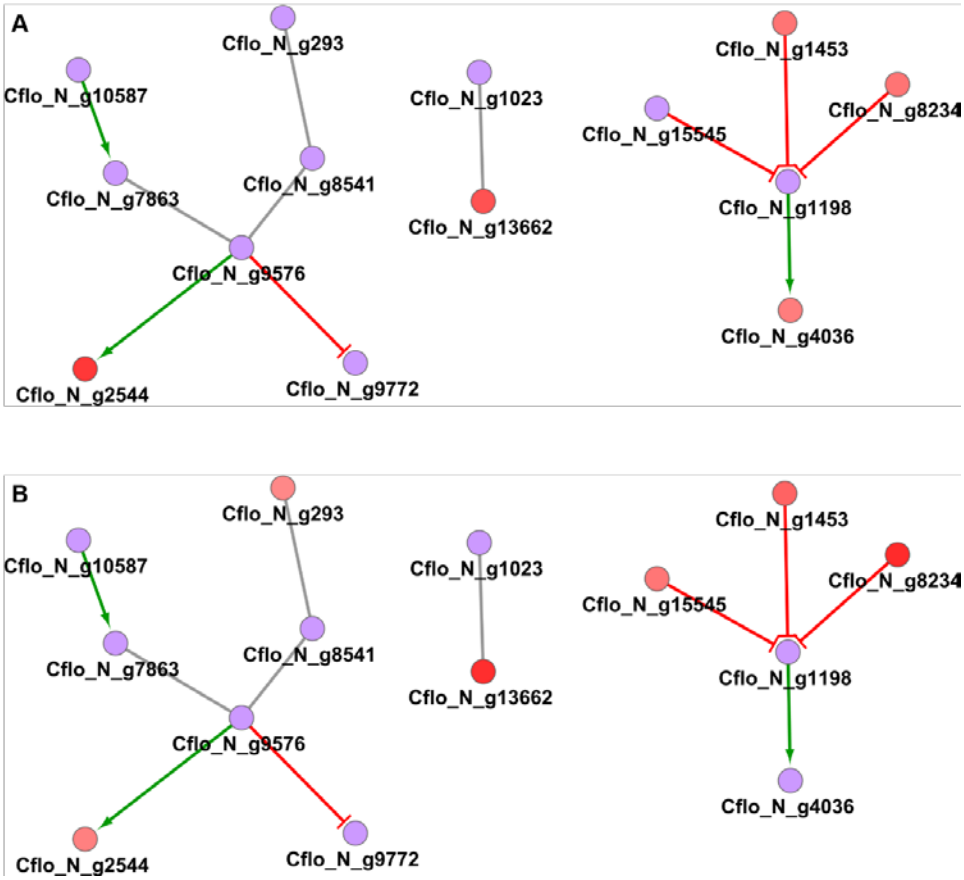


Figure 20. Conserved of active modules in *C. floridanus* identified from Illumina sequencing data and MS data. Node designation: red gradients, down-regulated proteins; green gradients, up-regulated proteins; violet, not differentially expressed. All nodes are denoted by re-annotation identifiers (see Appendix XI for annotation and expression value). Connectivity among nodes is based on undirected interactions (—). (A) Mapping of DESeq transcript expression data over conserved active modules. (B) Mapping of larvae haemolymph protein expression data over conserved active modules.

Multiple haemolymph related transcripts that were found to differentially expressed upon bacterial challenge were not identified to be differentially expressed in haemolymph protein expression analysis. This indicates the transcripts coding these genes might be in control of post transcriptional regulation. In total, 17 such differentially expressed transcripts from DESeq analysis were identified that interacts with miRNA (Fig. 21), and not differentially expressed in MS analysis.

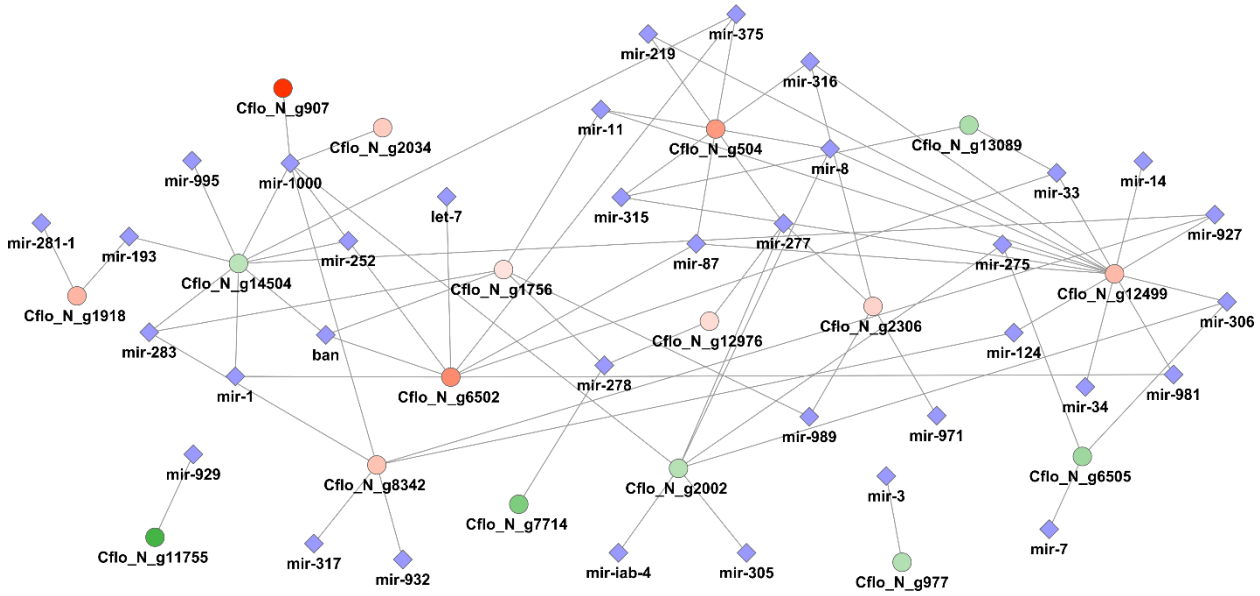


Figure 21. Interactions between miRNAs and differentially expressed haemolymph transcripts identified in Illumina data. Only the haemolymph transcripts are included here that were not significantly expressed at protein level in haemolymph. Node designation: red gradients, down-regulated proteins; green gradients, up-regulated proteins; blue, miRNAs. All nodes are denoted by re-annotation identifiers (see Appendix VI for annotation and expression value). Connectivity among nodes is based on undirected interactions (─).

4

Discussion

This chapter discuss the results of experiments and bioinformatics analyses. However, some results are also discussed together with results to increase the readability (specifically section 3.17 and 3.18). Some passages (specifically section 4.2 – 4.3) have been quoted *verbatim* from the published BMC Genomics article (Gupta et al., 2015) as I am further owner of the copyright of the article (licence type is CC BY 4.0).

4.1 Re-annotation improves the existing annotations

The annotation projects of large eukaryotic genomes often provide a broad but shallow view of the structure and function of genes as their fundamental perspective is to offer overview of entire genome rather than defining individual genes (Lewis et al., 2000). This subsequently leads to erroneous annotations of multiple genes in the complex eukaryotic genomes. However, the recent update in the annotation of *Rattus norvegicus* genome (Li et al., 2015) indicates the upcoming advancement in this component of genomics research. Here, based on transcriptome data re-annotation was performed of *C. floridanus* genome to improve the existing annotation. To achieve the improvement in structural annotation, creation of a good training set and integration of extrinsic evidences to gene predictors are key steps. In the upcoming text first the methods are briefly discussed then the achieved improvement in structural and functional annotations are discussed.

For any gene prediction machine learning approach, construction of training set of reliable gene structures is crucial. The degree of curation of gene structure annotation submitted to the various databases is often unclear (Munch and Krogh, 2006). Therefore, reliable training sets are hard to come for most organisms. To determine the genes with correct exon/intron structures multiple tests were performed, for instance, selecting the variable genes from different GO classes or random genes, and accuracy of the test set with the generated gene prediction parameters were tested. In addition, the tests were also performed by mixing

the training sets created from different tools. The comparative results of the accuracy of prediction by the training set created from various approaches provide the evidence that the training set created by the combination of methods such as Cegma (Parra et al., 2007), PASA2 (Haas et al., 2003) and Scipio (Keller et al., 2008) provide the high-quality training set. Gene models generated from the gene predictor trained with this training set provides the high base level and exon level accuracy. The parameters provided for training can be used for gene prediction in any other phylogenetically closer species, however training the predictor with the same species on which the user working on is always a good practice (Gupta et al., 2016).

The improved performance can be achieved by gene-predictors by integrating evidences for gene models. Using the Illumina sequencing, 244 million reads were generated and screened for the identification of reads showing full or partial alignment to the *C. floridanus* genome. Moreover, the ESTs and assembled reads were also used to correctly identify the gene features, for instance, better annotation of intron, exons and splicing sites. Such data allow the additional support for the accuracy of the gene models during the gene predictions. Moreover, using the genome with unmasked repeats could lead to over-predicted genes therefore, before running the pipeline all the repeats in the genome were identified and masked. Comparison of the improved genome annotation of *C. floridanus* to published data on other genomes revealed that *C. floridanus* genome has fewer repetitive elements than *D. melanogaster* (27.38 %) or *N. vitripennis* (24.31 %), but more than *A. mellifera* (6.86 %).

Even though re-annotation has a significant effect on data quality, the important question for knowledge generation is whether or not it has an impact on data interpretation. The impact of the presented re-annotation on the interpretation of this study can be described on three levels.

First, the accuracy of annotation for both structural and functional annotation was improved using the gene prediction tool Augustus also in terms of splicing events. Bonasio and co-workers (2011) detected 7583 alternative splicing events in 2538 genes (Cflo.v.3.3) overall (Bonasio et al., 2010), while the analysis presented here revealed 1928 genes affected by alternative splicing events coding for 4666 alternative transcripts. However, the Cflo.v.3.3 data available for *C. floridanus* contain 17,064 transcripts and 17,064 proteins without distinction between alternative isoforms and thus cannot be used for further analysis. The new data reported here can be accessed at web repository (http://camponotus.bioapps.biozentrum.uni-wuerzburg.de) and distinguish the alternative splicing products of the genes with the

suffix in the accession number as t1, t2, t3 etc. Furthermore, over-prediction (false positives) of single exon genes was also reduced in re-annotation. Similarly, the reduction of gene number and an increase in multi-exon genes was also observed in re-annotation effort of *Cucumis sativus* genome (Li et al., 2011). The prediction of single-exon genes is still unreliable in eukaryotic genomes, therefore the increase in transcript supported multi-exon genes indicates an improvement of the protein-coding prediction to some extent.

Second, the re-annotated data allowed the identification of gene models supported by evidences such as transcripts, exon and intron hints. Transcript evidence is taken as the most valuable evidence in protein-coding gene prediction, as it often capable to exactly identify intronic boundaries (Haas et al., 2002). RNA-Seq, among all the transcript evidence that affects gene prediction, is the one has distinctive ability for high coverage in protein-coding gene predictions, thus could increase the number of genes supported by transcript evidence and improve the structural annotations for both the single or multi-exon genes (Haas et al., 2002), for instance, during the re-annotation wrong annotation of PGRP-LC was spotted and subsequently corrected (Fig. 1), thus the propagation of this wrong annotation by Blast similarity search is somehow prevented by the presented transcriptome based re-annotation. Indeed, despite the great potential of RNA-Seq data for the annotation of protein-coding genes training gene finders on a newly assembled genome can be a challenging, frustrating task, so much that several genome projects attempt to leverage gene finders trained for other genomes (Law et al., 2015). Notably, sequence features as codon bias and splicing signals vary from organism to organism and even the nearest phylogenetic neighbor does not necessarily possess compatible parameters, therefore using the gene prediction parameters from other species has to compromise with the gene model accuracy (Holt and Yandell, 2011).

Third, with the re-annotated gene models, comparatively higher number of proteins were functionally annotated (see Fig. 2-5). Re-annotation identified additional genes involved in different pathways, specifically humoral immunity signalling pathways, that were not identified in the original work. Furthermore, functional domain containing multiple proteins that were labeled as hypothetical protein, their ortholog in re-annotation were named as ‘x’-domain containing protein, where ‘x’ is the name of domain. Thus, it also increased the coverage of annotating hypothetical proteins in re-annotated version. Using the hierarchal filtering system for determining the most informative functional annotation with AutoFACT, enabled the consistent naming of predicted proteins.

4.2 Comparative immunomics explores the *C. floridanus* immune system

To annotate the immune genes a repository of insect immune genes was established based on the extensive literature mining and curations. As defined earlier in this work, the insect consists the multilayered immune system with multiple processes to combat the pathogens and other harmful substances. To understand the infection process at system level, it is essential to establish the broad immune repertoire with better classification of genes and proteins involve in immunity. Unfortunately, even the GO classification do not classify well insect immune genes in all the different insect immune related categories and additionally, very little was known about the *C. floridanus* immune system. In order to establish the immune repertoire and present the classification of proteins in different components of *C. floridanus* immune system, homology based methods and GO classifications, followed by extensive manual curations were relied (Fig. 6). The immune related function of re-annotated proteins was assigned using orthology based function transfer from the previously annotated proteins in different insects. The fundamental principle of functional prediction of proteins relies on the evolutionary connections within protein families. Ortholog grouping can provide information regarding the evolutionary origin and functional conservation of proteins. The proteins clustered together are likely to be more closely related evolutionarily to each other than to proteins in other orthology groups (or singletons). Surprisingly, the presented immune repository of *C. floridanus* is quite bigger than *A. mellifera*, which further deviate the general impression that social insects are expected to have low number of immune proteins. The comparison of *C. floridanus* immune proteins with previous assembly of *A. mellifera*, orthologs of 186 *C. floridanus* immune proteins were identified however, comparison with re-assembled and re-annotated update of *A. mellifera* genome (Apis_4.5) revealed the presence of 374 immune proteins. This indicates the general statement for the reduction of immune genes is not completely applicable for all the social insects. Indeed, the social insects does not have large repository of AMPs, but probably this is compensated by the other immune factors. Notably, there might be possibility the genes are presented in the genome but not identified because of error in assembly or in annotation, for instance, in the update of *A. mellifera genome* ~5000 more proteins (Elsik et al., 2014) were identified than the previous assembly.

Although, the broad repository of *C. floridanus* immune genes were classified mainly based on orthology, GO annotation and literature there are other specific genes that could participate in immune response. Genes found to be differentially regulated by immune

challenge are likely to be involved in immune functions, although it is known that there are many regulatory factors overlaps with stress responses rather than immune challenge (Colgan et al., 2011). The signalling pathways involved in antimicrobial humoral immunity were constructed based on homology and literature. In accordance with previous findings (Sackton et al., 2013) signalling pathways seem to be more conserved in comparison to effector proteins that show higher levels of taxon specificity.

Besides PGRP-SD and DIF all the proteins of *D. melanogaster* Toll pathway were identified in *C. floridanus.* Though, the method used could not distinguish between GNBP1 and GNBP3 receptors. GNBP3 recognises fungal cell wall components while GNBP1 is known to perceive Lys-type PGN (Gottar et al., 2006). The previous expression data after immune challenge of *C. floridanus* with Gram-negative or Gram-positive bacteria revealed a long-lasting up-regulation of the gene encoding one of the GNBPs (EFN66519.1; Cflo_N_g5742t1) only after infection with Gram-positive bacteria (Ratzka et al., 2011). This suggests that this *C. floridanus* protein may be able to recognise Lys-type PGN and thus may be a functional homolog of GNBP1 of *D. melanogaster* (Fig. 8). Additionally, *C. floridanus* encodes a homolog of the protease Persephone which was previously shown to be involved in the detection of danger signals indicative for infection with Gram-positive bacteria and fungi (Ashok, 2009).

The role of Gram-negative bacteria triggered IMD pathway in *C. floridanus* is interesting as it involves in both the immunity and symbiosis. It has been shown that besides the stimulated antimicrobial activities, AMPs produced in response to the activated IMD cascade may also contribute in controlling the number of endosymbionts in *C. floridanus* midgut (Ratzka et al., 2011). In this pathway, misannotation of PGRP-LC protein was identified during the re-annotation procedure. PGRP-LC is extracellularly located pattern recognition receptor which was wrongly annotated as PGRP-LE in Cflo.v.3.3. This wrong annotation occurred probably because of the high similarity between PGRP-LC and PGRP-LE. In such cases, conclusions solely based on similarity, whether the annotation is correct or incorrect is problematic. Here, based on additional computational strategies such as orthologous clustering, domain analysis and gene bases phylogeny was used to confirm the error. Indeed, such errors are highly propagative as availability of this wrong annotation (EFN63542.1) in public databases may further mislead the scientific community for annotation of their protein of interest which is homolog of this sequence. Moreover, previously reported (Ratzka et al., 2013) two PGN recognition proteins with regulatory

function (PGRP-SC and PGRP-LB) were also identified in IMD pathway. Because of the presence of an amidase domain both of them may down-modulate the signalling pathways by cleavage of PGN (Bischoff et al., 2006; Royet and Dziarski, 2007). PGRP-LB was found to be involved in the tolerance towards the obligate intracellular endosymbiont *B.* floridanus in the midgut tissue during pupation of the *C. floridanus*. PRGP-LB is highly up-regulated only in the midgut tissue and not in other parts of the pupa body cavity, coinciding with a massive multiplication of the endosymbiont and a reduction of the immune competence in this tissue (Ashok, 2009). Homolog of one important component of *D. melanogaster* IMD pathway, key gene was not identified in *C. floridanus*. In *D. melanogaster*, Key (or Kenny) regulates the enzymatically active Ird5 subunit mediated phosphorylation of Relish and works as a complex. It has been published that a key mutant of *D. melanogaster* is highly susceptible to bacterial infections (Rutschmann et al., 2000). Iterative sequence analyses also verified the lack of the Key subunit in *A. mellifera*, *N. vitripennis* and other ant species, suggesting a common character of the IKK complex in hymenoptera. Whether the lack of the Key subunit may reflect a reduction in the immune potential of hymenoptera or whether so far unknown factors may be involved in building a functional IKK complex is not yet known. Moreover, most of the implicated core components of *D. melanogaster* JNK signalling pathway were identified to have apparent homologues in *C. floridanus* (Fig. 9).

Jak-Stat signalling pathway involves in wound healing, developmental processes, activation of stress-protective proteins, inflammatory and immune responses. Although, similar to *A. mellifera* the activators of cytokine receptor Dome were not identified in *C. floridanus* which indicates the activation of Jak-Stat cascade in *C. floridanus* is different from *D. melanogaster* (Fig. 10). TEPs are downstream effector molecules of Jak-Stat pathway which contributes to stem cell regulation in the intestine of *D. melanogaster* and thus to midgut homeostasis (Osman et al., 2012; Valanne, 2014) and in involves insect immunity by promoting phagocytosis of bacteria (Blandin and Levashina, 2004; Bou Aoun et al., 2010).

Interestingly, despite broad immune repository compare to *D. melanogaster* fewer AMPs could be annotated (Table 5). The apparently low number of AMPs may be compensated by the astonishing gene structure of hymenoptaecin which in the ants is encoded as a huge precursor protein with several repeated hymenoptaecin domains. Proteolytic maturation of this precursor protein leads to a massive amplification of the immune response (Ratzka et al., 2012). Additionally, the hymenoptaecin gene is among the most strongly induced genes after immune challenge. The apparently quite low number of AMPs and of

PGN recognising PRRs as described above may relate to the social lifestyle of ants and bees as compared to the solitary and parasitic lifestyle of the wasp *N. vitripennis*, since social life might allow hygienic measures on the colony level (Cremer et al., 2007). Moreover, ants produce a range of antimicrobial secretions that may be used to reduce pathogen pressure externally before an infection of the body occurs. Consequently, these external immune defence strategies may trade off against internal immune defences and may result in a reduction in the number of effector molecules (Otti et al., 2014; Vilcinskas, 2013).

4.3 RNA-Seq analysis identifies a comprehensive profile of DEGs

To identify the DEGs on bacterial infection in *C. floridanus*, the full transcriptome was analysed by Illumina sequencing (244 million reads), comparing challenged and unchallenged animals (pools of whole larvae L2 and workers W2 for each sample). CuffDiff and DESeq were both used to identify DEGs, and results were compared. Results from both tools showed high concordance ($R^2 = 0.9$). When considering the ontologies associated with the DEGs, the statistical overrepresentation of 44 GO terms (25 for biological process, 16 for molecular function and 3 for cellular components) relative to the rest of the whole proteome of *C. floridanus* were annotated (Table 6). Interestingly, the overrepresentation of oxidoreductase activity was found highly significant. In agreement with a recent study, the overrepresentation of oxidoreductase activity was also observed in mixed sample RNA-Seq analysis of carpenter ants *C. castaneus* and fungal pathogen *Ophiocordyceps unilateralis.* Oxidoreductases are known to produce ROS (Esterházy et al., 2008; Raha and Robinson, 2000), and also to control redox homeostasis (Messens et al., 2013). Therefore, it is conceivable that during immune challenge the activation reactive oxygen species (ROS) mediated immunity contribute to eliminate the pathogens and it might be one of the common defence strategies of carpenter ants against different microorganisms.

In the qRT-PCR experiments, it was observed that the induction of immune genes was much stronger in larvae than in workers. At first glance this seems somewhat surprising, since ant larvae are constantly cared for and groomed by nurse workers within the protected nest environment, while in particular foraging workers should be exposed more frequently and intensely to a feculent environment. However, larvae may be more vulnerable to pathogen infections due to their relatively thin and soft cuticle and the inability to groom themselves. Therefore, a highly responsive immune gene regulation in larvae may contribute to a long term colony success by ensuring a continuous supply of a large number of healthy offspring.

Indeed, recent infection experiments with *C. pennsylvanicus* larvae indicated that their individual immune response is important and brood care by nurses does not alleviate the individual immune competence of immature stages (Purcell and Chapuisat, 2014). Besides, brood care is also of prime importance since it was shown recently by cross-fostering experiments that in the ant *Formica selysi* the colony origin of care taker animals contributed to the resistance of freshly eclosed animals against an entomopathogenic fungus (Wilson-Rich et al., 2008). In *A. mellifera* it was also shown that developmental stages differ in immunocompetence – larvae and pupae had the highest haemocyte counts while adult workers had the strongest phenoloxidase activity (Richard et al., *2012).*

Overall, in humoral signalling pathways very few transcripts were found to be differentially expressed but the differential expression of downstream molecules (AMPs and TEPs), TFs and receptors of Toll and IMD pathways were identified. This indicates, there might be unconventional signalling to activates the effectors of these pathways, for instance, the activation of effectors by direct targeting of pathogen or the unknown mediators of the *C. floridanus* immune system. Furthermore, the activation of the components of cellular immunity were also observed which indicates the cellular immunity could play the important role in *C. floridanus* to combat with infection. The analysis of interactome further strengthen these facts.

4.4 Network analysis identifies important proteins implicated in immune response

PPI and signalling network for *C. floridanus* were constructed using the interolog and signalog method based on the orthologous genes interactions and signalling protein interactions respectively. The combination of methods could improve the assignment of functional orthologs (Pereira et al., 2014), hence to improve the quality of the orthology based prediction two leading orthology prediction methods InParanoid and OrthoMCL were used to reconstruct preliminary interactome. Notably, these methods can include the outparalogs in the orthology groups which further may impact the correct prediction of PPI network therefore, only the high confidence seed orthologs and the consensus prediction of both the selected tools were considered. For a reconstructed interactome, a dataset of high coverage with more false positive is not much valuable hence the improvement of the reliability of PPIs is critical before analyzing the PPI networks.

To avoid the over-prediction several filters were applied for to reconstruct high confidence interactome. Only the interactions having possibility to interact at structural level were selected based of occurrence of experimentally verified DDI data. Considering the accuracy of the PPI map, adding the DDI information in PPIs is one of the powerful methods to minimize the number of false positives (Wojcik and Schachter, 2001; Zhang et al., 2009; Zhou et al., 2013). Afterwards such interactions were removed where the interacting proteins had apparently different localization and in case of multiple localizations they do not share any of their localization compartments. During the localization prediction the inconstancy were observed in the prediction from different tools. Therefore, to deal with the different results of localization from different sources the agreement of localization prediction with Swiss-Prot annotations were considered. The comparison of six tools (WolfPsort, Cello2.5, LocTree3, UniLoc, YLoc2 and ILoc-Euk) for single localization ant proteins in DDI filtered interactome (in three compartments; nucleus, cytoplasm or mitochondria) showed the comparatively higher consistency with UniLoc and Swiss-Prot predictions (Appendix XII). Hence, the prediction from UniLoc server and Swiss-Prot were considered as reliable for *C. floridanus* proteins. Finally, isoforms were represented by single node to reduce the noise in data. These filters subsequently reduce the interactions and nodes but finally results in high confidence network. Although, the interolog based PPI approaches are well-established however, they rely on template interactions and unfortunately so far in most HTP experiments extracellular, membrane-bound and nuclear proteins are underrepresented which affects the coverage of interactome and usability for identifying signalling interactions. Another limitation of HTP PPI detection assays is that they produce undirected interactions, though in signalling directions are essential. Accordingly, several signalling pathway databases have been created recently by manually collecting the directed interactions from the literature (Bauer-Mehren et al., 2009). Using different sources of signalling interactions in *D. melanogaster*, the signalome of *C. floridanus* was also constructed and ultimately merged with *C. floridanus* interactome to derive the IISC network which consist both the directed and undirected interactions. The considered IISC network, despite its shortcomings in data coverage, provides an excellent framework for navigating through the *C. floridanus* proteome. It also allows for refinement of the network upon the availability of new experimental data.

Generally, the highly connected proteins in the interactome contribute to the functional integrity of the network and from the topological perspective; the removal of such

critical nodes renders the network collapsed into isolated clusters. The degree of the node (Batada et al., 2006) and the betweenness centrality (Yu et al., 2007) represent the most important network features as the high-degree nodes (the hubs) and the nodes with high-betweenness centrality (the bottlenecks) plays a central role in maintaining the network connectivity and functional integrity. The mapping of expression data over the reconstructed networks provides the valuable information to understand the key players involve in immunity at systems level, for instance, differentially expressed Notch (Cflo_N_g6260) identified as top hub among all DEGs with 37 interactors which is a multifunctional and could affect multiple processes during immune responses (see Results). The networks presented here can be further used in different studies.

4.5 *S. marcescens* can directly interfere with the different layer of *C. floridanus* immune system

The results of host-pathogen interactions indicate that membrane, secretory and virulence associated proteins of *Serratia* can interact with the proteins present in different layer of host immune system. The relevance of interesting interactions is discussed together with the results. In summary, hubs were identified as preferential targets. Additionally, the targeting of humoral immunity pathway indicates the possibilities of stimulation of these pathway can also occur by non-conventional approach. The Illumina sequencing data analysis showed in these pathways mostly the top receptors, inhibitors, and downstream TF were differentially expressed while the signalling components in the middle of the pathway did not show significant change in expression after immune challenge. Such results can serve as testable hypotheses especially relating to host immunity for which focused experimentation might increase our understanding of patterns of host-pathogen interactions.

4.6 *B. floridanus* can interact with *C. floridanus* immune system

Immune systems play a central role in the way insects distinguish between pathogenic and symbiotic bacteria and regulate their immune response accordingly. Symbiotic bacteria are important in insect hosts, but the interaction studies of insect host and endosymbionts have been largely overlooked as they have proved difficult to culture in the laboratory. To overcome this, computational approach of interspecies PPIs prediction was used to establish the first draft of *C. floridanus* - *B. floridanus* interaction network. Interestingly, most of the *B. floridanus* proteins identified to interact with host also had ortholog in *S. marcescens*. Moreover, as the results shows 32 interactions were found to be highly conserved in five other

insect-endosymbiont pairs. In the highly conserved PPIs several interactions were identified that could be subject of future experimental validation, for instance, host collagen alpha-1(IV) (Cflo_N_g1734) was found to be in interaction with endosymbiont alkyl hydroperoxide reductase C22 protein (AhpC). The role of collagen alpha-1(IV) has been reported in insect immunity (Altincicek and Vilcinskas, 2006) while the transcriptional profiling of the *B. floridanus* has shown high transcript level of ahpC gene during different developmental stages of *C. floridanus* (Stoll et al., 2009). Studies of the such evolutionary conserved insects-endosymbiont interactions could yield valuable information about how bacteria infect host cells, avoid immune responses, and manipulate host physiology.

4.7 Insights from the haemolymph protein expression

Proteins generally are of higher abundance and have a longer half-life within cells than do transcripts therefore, a more accurate reflection of biological function may be gained from the proteome. Besides, many proteins are post-translationally modified, which cannot be detected in the transcriptome. It has been estimated that 80 % of a cell's phenotype can be defined by the proteome compared with only 30 % by the transcriptome (Feder and Walser, 2005). To explore the protein expression upon immune challenge, the haemolymph of infected and control larvae and worker ants were subjected to MS/MS analysis.

The comparison of top GO annotation of differentially expressed proteins/genes suggested the differences in the immune responses of larvae and adults. In the Illumina data highly significant overrepresented GO categories includes oxidoreductase activity and hydrolase activity. Moreover, the larvae MS data enriched hydrolase activity while in adults, oxidoreductase activity was enriched under highly significant overrepresentation of GO categories. This suggests the larvae mostly rely on protease and hydrolase activities rather than ROS-mediated immunity which could be a dominant process in adults to combat the pathogens.

The mapping of protein expression data over IISC network disclosed the highly connected subnetwork in larvae immune response in compare to adults where very few active modules could be identified. Furthermore, in agreement with the previous findings (Ratzka et al., 2011) the conserved active module between the larvae proteome and transcriptome data again highlighted the role of Lysozyme c-1 in *C. floridanus* immunity.

Notably, a Grp7_allergen domain containing hypothetical protein (Cflo_N_g7901t1) showed downregulation both at protein and mRNA level. CATH evaluation of this protein

revealed the presence of CATH domain 1ewfA01 which represents the bactericidal permeability-increasing (BPI) protein like folds. BPI has been known for decades to bind with LPS (lipopolysaccharide), thus involves in host defence against Gram-negative bacteria (Elsbach, 1994). The early effects of BPI on Gram-negative bacteria are synergistically enhanced by defensins and other AMPs (Elsbach et al., 1994). The strong affinity of BPI, to interact with Gram-negative bacteria is attributed by its very basic N-terminal half that helps in destruction of negatively charged LPS. However, the antibacterial activity of BPI is inhibited by LPS from opportunistic bacteria *P. aeruginosa* (Wasiluk et al., 1991). Other ants also showed the presence of homolog of this allergen which typically involve in inflammatory and hypersensitive reactions but the differential expression, presence of BPI and also the structure similarity of Grp7_allergen to different Toll pathway components provide the evidence of connections between allergic reaction and innate immunity (Mueller et al., 2010).

The re-annotation suggested the *C. floridanus* genome encodes for MD-2-related lipid-recognition domain containing three Niemann-Pick C2 (NPC2) like proteins. Interestingly, the downregulation of one NPC2 (Cflo_N_g1756t1) was identified in Illumina sequencing data while one other NPC2 (Cflo_N_g7354t1) was found up-regulated in larvae haemolymph in MS analysis. NPC2 can induce the IMD pathway by recognizing the lipid product from bacteria (Shi et al., 2012) and further the differential expression of NPC2 was also reported in immune-stimulated honeybees (Richard et al., 2012), which indicates the importance of NPC2 in *C. floridanus* immunity.

Interestingly, downregulation of a novel AMP waprin-Phi1 (Cflo_N_g4956t1) was identified in adults MS data while the DESeq analysis showed the significant increase in sequenced reads of waprin-Phi1 after immune challenging at mRNA level. However, it was not reported in the transcriptome study as the adjusted p-value for the differential expression of this transcript was 0.057, and in the study the differentially expressed transcript strictly chosen at adjusted p-value threshold ≤ 0.05. The Blast search showed the presence of similar waprin-like proteins in several hymenopterans species including ants. Similar to crustin IV AMPs of crustaceans, waprin-Phi1 consist two whey acidic protein (WAP) domains.

Double domain WAPs have never been reported from ant however, the single domain WAPs has been reported in venom transcriptome of ant *Tetramorium bicarinatum* (Bouzid et al., 2014). Waprins were firstly identified in spitting cobra *Naja nigricollis* (Torres et al., 2003). Besides reptiles, the WAP domain containing proteins were also identified in

amphibians and crustaceans (Smith, 2011). In crustaceans, WAP family AMPs are referred as crustins. Comparative genomics of seven ant APMs reveals the presence of one crustin V like AMP encoded by *C. floridanus* which contains the cystin rich region, an aromatic residue rich core and the signature WAP domain (Zhang and Zhu, 2012). The AMP reported here is different from crustin V, and like crustin IV it consists a signal peptide and two WAP domains. The double domain WAPs family has been described from many tissue types including haemocytes, and different organisms and has been found to have many functions, including, immune-modulation, peptidase inhibitor activities and antimicrobial responses (Du et al., 2009; Li et al., 2013; Liu et al., 2013).

5

Conclusion

In this thesis annotation of *Camponotus* transcriptome provided for the first time not only a transcriptome-based update of the genome annotation, but also a description of mRNA, splicing and transcriptome in *C. floridanus* as well as a detailed study on gene expression (challenged/unchallenged) of immune response in *C. floridanus* (Gupta et al., 2015). In particular, the analysis allowed to extend the previously annotated protein repertoire not only by about 20% (including splicing variants), but was instrumental in analyzing the immune response of *C. floridanus* and to newly identify nine putative *Camponotus*-specific proteins possibly involved in immune functions or stress response. Interestingly, a new AMP waprin-Phi1 was identified first time in *C. floridanus*. The gene expression analysis suggests significant stage specific differences in the immune responsiveness of *C. floridanus* which may be an important feature possibly contributing to colony success, requiring, however, further investigation in the future. Similar observation of stage specific differences in immune response were also identified by analysis of mass-spectrometry data of (challenged/unchallenged) *C. floridanus* larvae and workers haemolymph (unpublished results).

Analysis of reconstructed PPI and signalling networks after integrating them with transcriptome and proteome data offered several general conclusions (for instance, (i) Notch plays an important role in interactions and signal transduction during response against bacteria, especially it participates in the pathogen-induced epithelial renewal program (Jiang et al., 2009). (ii) *Serratia* interacts with several hub proteins and signalling components in multiple layers of *C. floridanus* immune system. (iii) *B. floridanus* could interfere with the *C. floridanus* immune system by interacting with the host proteins. (iv) Induction of Lysozyme c-1 by the signals originated from destruction of cuticles plays in important role is *C. floridanus* immunity. (v) Larvae seems to fight against infection in much coordinated way

while workers might adopt the strong ROS mediated immunity as suggested by qPCR, network and over-representation analysis) (unpublished results).

Considering that *C. floridanus* belongs to the most prominent ant family *Camponotus* and that ants are the most successful insect order, comprising up to 25% of the biomass in a tropical forest this is an important contribution for understanding insect biology and in particular the insect immune system. *C. floridanus* needs the endosymbiont *Blochmannia* and a challenge of its immunity is to fight related Gram-negative bacterial species such as *S. marcescen*s while keeping the endosymbiont. Surprisingly it was identified that the immune system of *Camponotus* ants is equally rich as in other insects, including diverse antimicrobial peptides and does not rely more on 'social immunity' as other insects. However, more detailed investigations, in particular functional studies have now to be based on this first transcriptomics-driven overview to understand the immune response better, in part hampered by a lack of genetic tools in *Camponotus*. A number of other related publications appeared (see page 114) in addition to (Gupta et al., 2015; Gupta et al., 2016), including four first or shared first author publications.

The methods used include a refinement of splicing prediction in the well-known software package Augustus and a whole pipeline of scripts and methods to identify orthologues and filter proteins by different criteria. This has lead also to two further shared first or first author publications in several journals characterizing drug pipelines (see page 114). The method proposed for re-annotation is published as book chapter (Gupta et al., 2016). Further publications are in pipe, for instance the characterization of the immune proteome in *Camponotus* by an overview of the haemolymph, the study of *C. floridanus* interactome, signalome, host-pathogen and host-symbiont PPI interactions.

Taken together, by refining the accuracy of prediction regarding single and multiple transcript genes and evidence for their expression the transcriptome sequencing improved the annotation of the *C. floridanus* proteome, the identification of repetitive elements as well as alternative splicing predictions. Moreover, the presented ant interactome and signalome from this study is the first large-scale PPI network of any sequenced ant will serve for more comprehensive screening of cellular operations during pathogen attack. The ant-bacteria interspecies interactomes will contribute to identifying bacterial infections and tolerance mechanisms in insects, detecting important target proteins and thus promoting potential biocontrol innovations against agricultural insect pests.

Appendix I. Functional annotation of highly conserved immune proteins shared by *D. melanogaster*, *A. mellifera*, *N. vitripennis* and *C. floridanus*.

Accession Number	Function/Immune pathways	Annotation
Cflo_N_g11181t1	Coagulation, Hematopoiesis, Encapsulatin/Nodulation	Hemocytin
Cflo_N_g1329t1	crq-family scavanger receptor, Microbial recognition	Scavenger receptor class B member 1 (Fragment)
Cflo_N_g14772t1	crq-family scavanger receptor, Microbial recognition	scavenger receptor class B member 1-like
Cflo_N_g15204t1	crq-family scavanger receptor, Microbial recognition	Scavenger receptor class B member 1
Cflo_N_g9951t1	crq-family scavanger receptor, Microbial recognition	Scavenger receptor class B member 1
Cflo_N_g6152t1	Draper, Microbial recognition, Haemocyte receptor, Phagocytosis, Laminin	Multiple epidermal growth factor-like domains 10
Cflo_N_g9696t1	Eater, Haemocyte receptor, Microbial recognition, Phagocytosis	von Willebrand factor D and EGF domain-containing protein
Cflo_N_g9695t1	EGF, Haemocyte receptor, Microbial recognition, Phagocytosis	von Willebrand factor D and EGF domain-containing protein
Cflo_N_g2961t1	Encapsulation lectin, Galectin, Microbial recognition	Macrophage mannose receptor 1 (Fragment)
Cflo_N_g5742t1	GNBP, Microbial recognition, Beta-1,3-glucan-binding protein, Toll pathway	Beta-1,3-glucan-binding protein (GNBP)
Cflo_N_g15215t1	GNBP, Toll pathway, Microbial recognition	Beta-1,3-glucan-binding protein (GNBP)
Cflo_N_g12197t1	Haemocyte receptor, crq-family scavanger receptor, Phagocytosis, Microbial recognition	Protein croquemort
Cflo_N_g6595t1	Haemocyte receptor, crq-family scavanger receptor, Phagocytosis, Microbial recognition	Protein croquemort
Cflo_N_g79t1	Haemocyte receptor, Scavanger family, Phagocytosis, Microbial recognition	MAM domain-containing glycosylphosphatidylinositol anchor protein 1
Cflo_N_g79t2	Haemocyte receptor, Scavanger family, Phagocytosis, Microbial recognition	MAM domain-containing glycosylphosphatidylinositol anchor protein 1
Cflo_N_g10553t1	IMD pathway	Ubiquitin-conjugating enzyme E2-17 kDa, putative
Cflo_N_g10862t1	IMD pathway	Death domain-containing adapter protein BG4
Cflo_N_g2123t1	IMD pathway	Apoptosis 2 inhibitor
Cflo_N_g5881t1	IMD pathway	Inhibitor of NF-κB kinase subunit beta
Cflo_N_g7081t1	IMD pathway	Receptor-interacting serine/threonine-protein kinase 1
Cflo_N_g9451t1	IMD pathway	Mitogen-activated protein kinase kinase kinase 7
Cflo_N_g9451t2	IMD pathway	Mitogen-activated protein kinase kinase kinase 7
Cflo_N_g11129t1	Jak-Stat pathway	Suppressor of cytokine signalling 5
Cflo_N_g13035t1	Jak-Stat pathway	Signal transducer and activator of transcription
Cflo_N_g4220t1	Jak-Stat pathway	tyrosine-protein kinase hopscotch isoform X4
Cflo_N_g6115t1	Jak-Stat pathway	Cytokine receptor
Cflo_N_g15516t1	JNK pathway	Dual specificity mitogen-activated protein kinase kinase 7
Cflo_N_g3291t1	JNK pathway	Transcription factor AP-1
Cflo_N_g647t1	JNK pathway	Dual specificity mitogen-activated protein kinase kinase 7
Cflo_N_g647t2	JNK pathway	Dual specificity mitogen-activated protein kinase kinase 7

Cflo_N_g6920t1	JNK pathway, MAPK pathway	stress-activated protein kinase JNK
Cflo_N_g6920t2	JNK pathway, MAPK pathway	Stress-activated protein kinase JNK
Cflo_N_g11714t1	JNK pathway, Rho GTPase cytoskeleton	Ras-related protein Rac1
Cflo_N_g4036t1	Lysozyme	Lysozyme c-1
Cflo_N_g5519t1	Lysozyme	Lysozyme c-1
Cflo_N_g1918t1	Melanisation proPO	Phenoloxidase subunit A3
Cflo_N_g10272t1	PGRP, IMD pathway, Microbial recognition	PGRP-LC
Cflo_N_g102t2	PGRP, IMD pathway, Microbial recognition	PGRP-SC2
Cflo_N_g102t1	PGRP, Microbial recognition	PGRP-SC2
Cflo_N_g8526t1	PGRP, Microbial recognition, Toll pathway	PGRP2 (=PGRP-SA)
Cflo_N_g14922t1	RNA-I antiviral defence, Toll pathway	Insulin-like growth factor-binding protein complex acid labile chain
Cflo_N_g5858t1	RNA-I antiviral defence, Toll pathway	Protein toll (Fragment)
Cflo_N_g6513t1	RNA-I antiviral defence, Toll pathway	Protein toll
Cflo_N_g8274t1	RNA-I antiviral defence, Toll pathway	Protein toll
Cflo_N_g8278t1	RNA-I antiviral defence, Toll pathway	Protein toll
Cflo_N_g13088t1	Seprin, Melanisation, Toll pathway	Serpin B10
Cflo_N_g13089t1	Seprin, Melanisation, Toll pathway	Serpin B10
Cflo_N_g13089t2	Seprin, Melanisation, Toll pathway	Serpin B10
Cflo_N_g13089t3	Seprin, Melanisation, Toll pathway	Serpin B10
Cflo_N_g10213t1	Serine protease, Melanisation, Toll pathway	Thiamine transporter 2
Cflo_N_g7438t1	Serine protease, Melanisation, Toll pathway	Coagulation factor IX (Fragment)
Cflo_N_g8442t1	Serine protease, Melanisation, Toll pathway	Serine protease persephone
Cflo_N_g8446t1	Serine protease, Melanisation, Toll pathway	Serine protease persephone
Cflo_N_g9525t2	Serine protease, Melanisation, Toll pathway	Serine protease easter 2
Cflo_N_g9745t1	TEPs, Phagocytosis	CD109 antigen
Cflo_N_g4492t1	TEPs, Phagocytosis, Microbial recognition	CD109 antigen
Cflo_N_g7345t1	TEPs, Phagocytosis, Microbial recognition	Alpha-2-macroglobulin
Cflo_N_g9745t2	TEPs, Phagocytosis, Microbial recognition	CD109 antigen (Fragment)
Cflo_N_g11593t1	Toll pathway	Myeloid differentiation primary response protein MyD88
Cflo_N_g12735t1	Toll pathway	Protein spaetzle
Cflo_N_g1330t1	Toll pathway	Serine/threonine-protein kinase pelle
Cflo_N_g14413t1	Toll pathway	NF-κB inhibitor cactus
Cflo_N_g14414t1	Toll pathway	stress-induced-phosphoprotein 1-like
Cflo_N_g3305t1	Toll pathway	Embryonic polarity protein dorsal
Cflo_N_g9743t1	Toll pathway	Protein toll

Appendix II. List of serine proteases and serine protease inhibitors of *C. floridanus*. Abbreviation: 'dw' – down-regulated, 'up' – up-regulated and 'sc' – scaffold.

Sequence id	AutoFACT annotation	Immune regulation	Locus	Orientation	Transcript
Cflo_N_g9524t1	Serine protease easter 1	-	sc994	-	57578-60693
Cflo_N_g9525t2	Serine protease easter 2	-	sc994	-	64730-67689
Cflo_N_g9526t1	Serine protease easter	-	sc994	-	70512-75870
Cflo_N_g5965t1	Serine protease HTRA2, mitochondrial	-	sc518	+	12438-18006
Cflo_N_g8442t1	Serine protease persephone	up	sc816	+	98753-102340
Cflo_N_g8446t1	Serine protease persephone	-	sc816	+	112713-116259
Cflo_N_g7945t1	Serine protease snake 2	-	sc715	-	170525-173170
Cflo_N_g8441t1	Serine protease snake 1	-	sc816	-	93562-98632
Cflo_N_g10821t1	Putative serine protease K12H4.7	-	sc1292	-	14807-20708
Cflo_N_g15264t1	Mannan-binding lectin serine protease 1	-	sc6195	-	65-571
Cflo_N_g5224t1	putative serine protease K12H4.7-like	-	sc407	+	721941-723954
Cflo_N_g9525t1	serine protease easter-like	-	sc994	-	64730-64945
Cflo_N_g928t1	serine protease gd-like	-	sc64	+	1398654-1401280
Cflo_N_g929t1	serine protease gd-like	-	sc64	+	1405534-1411081
Cflo_N_g931t1	serine protease gd-like	-	sc64	+	1421825-1427704
Cflo_N_g4362t1	venom serine protease 34 isoform X1	-	sc383	+	428227-431257
Cflo_N_g4365t1	venom serine protease 34-like	dw	sc383	-	508255-510651
Cflo_N_g14393t1	serine proteinase stubble	-	sc3258	+	120-1098
Cflo_N_g2231t1	Serine proteinase stubble	-	sc200	+	140457-144635
Cflo_N_g2977t1	Serine proteinase stubble	-	sc255	-	36212-44566
Cflo_N_g2977t4	Serine proteinase stubble	-	sc255	-	36212-49622
Cflo_N_g12631t2	Serine proteinase stubble	up	sc1656	+	125268-136670
Cflo_N_g3399t1	Serine proteinase stubble	-	sc331	-	70277-70699
Cflo_N_g3400t1	Serine proteinase stubble	-	sc331	-	70806-77213
Cflo_N_g4559t1	Serine proteinase stubble	-	sc408	+	148213-171842
Cflo_N_g14403t1	Serine proteinase stubble	up	sc3265	-	11452-13554
Cflo_N_g9592t1	Serine proteinase stubble	-	sc968	-	404884-429821
Cflo_N_g9592t2	Serine proteinase stubble	-	sc968	-	404884-429821
Cflo_N_g12631t1	Serine proteinase stubble	-	sc1656	+	136302-136670
Cflo_N_g148t1	Serine proteinase stubble	-	sc45	-	1539-1914
Cflo_N_g7691t1	Serine proteinase stubble	up	sc704	-	2039-2617
Cflo_N_g12982t1	Serine proteinase stubble	-	sc2066	-	306-5229
Cflo_N_g15522t1	Serine proteinase stubble (Fragment)	-	C3970447	+	216-665
Cflo_N_g4557t1	Serine proteinase stubble (Fragment)	-	sc408	+	74925-105942
Serine protease inhibitors					
Cflo_N_g13088t1	Serpin B10	-	sc1993	-	53125-54393

Cflo_N_g13089t1	Serpin B10	-	sc1993	+	54890-76158
Cflo_N_g13089t2	Serpin B10	-	sc1993	+	54890-65924
Cflo_N_g13089t3	Serpin B10	up	sc1993	+	54890-58658
Cflo_N_g13647t2	Serpin I2	-	sc2479	-	1513-5256
Cflo_N_g1319t1	serine protease inhibitor 3-like	dw	sc57	+	236728-237415
Cflo_N_g2442t1	Kazal-type serine protease inhibitor domain-containing protein 1	-	sc107	+	779249-784824
Cflo_N_g3252t1	Serine protease inhibitor dipetalogastin	up	sc222	-	106326-110168
Cflo_N_g7717t1	Kazal-type proteinase inhibitor-like protein (Fragment)	-	sc704	+	671602-673222
Cflo_N_g4516t1	Kazal-type serine protease inhibitor	up	sc316	+	1792888-1795607

Appendix III. List of putative chitinases of *C. floridanus* and distribution of conserved motifs in their deduced amino acid sequences.

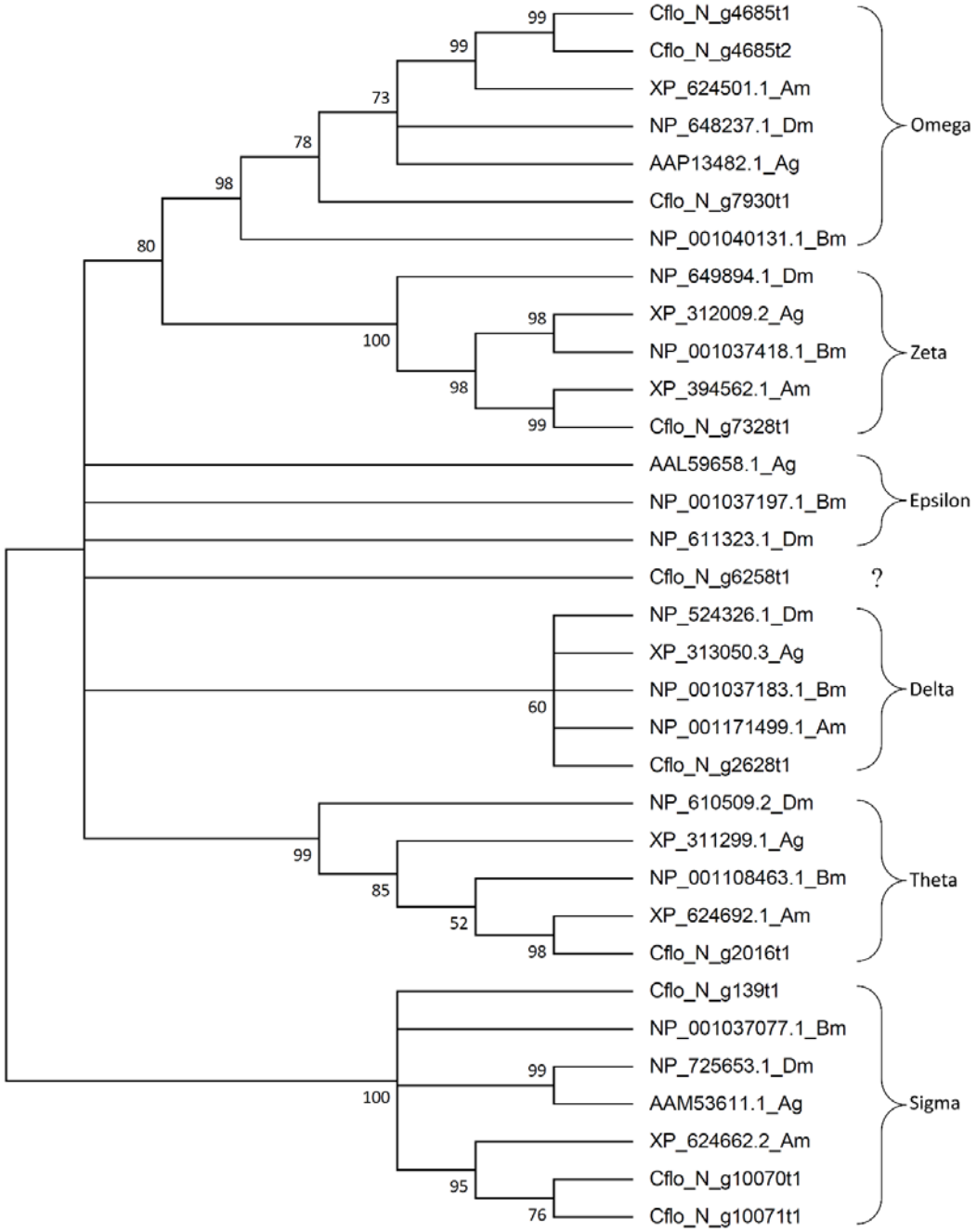
Accession no.	Position of conserved motifs			
	Motif I KXXXXXGGW	Motif II FDGXDLDWEYP	Motif III MXYDXXG	Motif IV GXXXWXXDXD
Cflo_N_g2277t1	82-90	NF	194-200	NF
Cflo_N_g8995t1	245-253	NF	357-363	491-500
Cflo_N_g14547t1	NF	NF	NF	NF
Cflo_N_g10513t1	1190-1198, 1601-1609, 2122-2130	1232-1242	508-514, 1304-1310, 1719-1725	188-197, 648-657, 1450-1459, 1865-1874, 2391-2400
Cflo_N_g10891t1	156-164, 569-577	NF	259-265, 684-690	405-414, 831-840
Cflo_N_g260t1	NF	NF	NF	NF
Cflo_N_g12931t1	101-109	142-152	217-223	370-379
Cflo_N_g8158t1	59-67	100-110	NF	330-339
Cflo_N_g8158t2	59-67	100-110	NF	330-339
Cflo_N_g7573t1	NF	NF	NF	NF
Cflo_N_g9837t1	NF	NF	NF	220-229
Cflo_N_g9838t1	NF	NF	NF	NF
Cflo_N_g10512t1	NF	NF	NF	NF

Note – Gene Cflo_N_g8158 codes two alternative spliced protein products Cflo_N_g8158t1 and Cflo_N_g8158t2. Abbreviation NF – not found.

Appendix IV. Number of chitinases, glutathione S-transferases, nitric oxide synthases, thioester-containing proteins, lysozymes and prophenoloxidases encoded by *C. floridanus*, other ants, *A. mellifera*, *N. vitripennis* and *D. melanogaster*.

Organism	Chitinases	GSTs	NOS	TEPs	Lysozymes	Phenoloxidases
Camponotus floridanus	13	10	1	4	3	1
Atta cephalotes	10	11	2	4	3	1
Acromyrmex echinatior	11	11	1	3	3	1
Pogonomyrmex barbatus	10	10	1	3	2	1
Harpegnathos saltator	12	9	1	3	2	1
Linepithema humile	9	8	1	3	2	1
Solenopsis invicta	9	9	1	4	3	1
Cerapachys biroi	9	10	1	3	2	2
Apis mellifera	9	10	1	3	3	1
Nasonia vitripennis	13	11	2	3	2	3
Drosophila melanogaster	8	20	1	5	7	3

Appendix V. Phylogenetic relationship between GSTs of *C. floridanus* and of other insects as inferred using the Neighbor Joining algorithm. The statistical reliability of the phylogenetic tree was tested by bootstrap analyses with 10,000 replications. The topology is based on a 50 % condensed tree obtained by bootstrap analysis. The percentage of replicate trees in which the associated nodes clustered together in the bootstrap test is shown next to the branches. Species abbreviations occur after the GenBank accession numbers are as follows: Dm = *Drosophila melanogaster*, Ag = *Anopheles gambiae*, Am = *Apis mellifera*, Bm = *Bombyx mori*, Aa = *Aedes aegypti*, Cf = *Camponotus floridanus*.



Appendix VI. Complete list of genes of *C. floridanus* which are differentially expressed after immune challenge. The genes listed were retrieved based on the analysis of the Illumina sequencing data using the programs Cufflinks and DESeq. The table shows significantly up-regulated and down-regulated genes 12 h after picking of larvae and workers with a 1:1 mix of Gram-negative and Gram-positive bacteria.

Re-annotation ID	Annotation	Cufflinks log2 (fold change)	Cufflinks q-value	DESeq log2 (fold change)	DESeq p-value adjusted
Cflo_N_g2215	Putative chitin binding peritrophin-a domain protein (Fragment)	-8.0072	0	-7.9683	5.00E-53
Cflo_N_g1319	serine protease inhibitor 3-like	-6.9677	6.98E-11	-7.4057	1.59E-50
Cflo_N_g907	Chymotrypsin-1	-6.1752	0.00083	-6.2845	3.81E-41
Cflo_N_g6949	Chitin_bind_3 domain containing protein	-5.2194	4.34E-05	-6.2105	6.30E-31
Cflo_N_g2154	Chymotrypsin-2 (Fragment)	-6.1153	3.75E-13	-6.1890	4.32E-40
Cflo_N_g4308	Lipase member H	-6.4288	3.75E-13	-6.1459	2.84E-25
Cflo_N_g10280	Pancreatic triacylglycerol lipase (Fragment)	-5.5895	0.99998	-5.5740	1.52E-33
Cflo_N_g5992	Hypothetical protein	-5.2456	6.86E-07	-5.2064	9.05E-07
Cflo_N_g12360	Hypothetical protein	0.0000	1	-4.8993	1.77E-28
Cflo_N_g6859	Flexible cuticle protein 12	-4.5719	0	-4.7130	1.95E-26
Cflo_N_g8235	Hypothetical protein	-3.8791	1.52E-05	-4.5981	0.00036
Cflo_N_g5438	Hypothetical protein	-2.8237	0.11376	-4.5625	0.00467
Cflo_N_g4736	Probable guanine deaminase	-4.7389	2.57E-10	-4.5600	8.65E-19
Cflo_N_g2249	AGAP011046-PA (Fragment)	-4.0283	0.00099	-4.1065	9.71E-21
Cflo_N_g6502	Arylphorin subunit alpha	-3.6604	3.04E-09	-3.7490	3.62E-18
Cflo_N_g922	A disintegrin and metalloproteinase with thrombospondin motifs 1	-3.6577	1.34E-09	-3.6982	1.60E-17
Cflo_N_g2544	Mite allergen Der f 3/Trypsin-1	-3.4496	5.54E-06	-3.5981	4.68E-17
Cflo_N_g12358	Hypothetical protein	-3.3663	2.11E-10	-3.3795	1.79E-13
Cflo_N_g4569	Sugar transporter ERD6-like 6	-3.2711	5.86E-07	-3.3537	9.75E-10
Cflo_N_g504	Carbonic anhydrase 2	-3.2820	0.01182	-3.3467	5.32E-15
Cflo_N_g14742	Guanine deaminase	-3.2502	2.36E-06	-3.3155	2.25E-11
Cflo_N_g3591	Transcription factor MafA	-3.3665	1.17E-06	-3.2858	1.68E-10
Cflo_N_g12015	Endocuticle structural glycoprotein SgAbd-1 (Fragment)	-3.1764	9.25E-05	-3.2723	2.02E-14
Cflo_N_g3206	cuticle protein 18.7-like	-3.0675	1.92E-05	-3.2544	7.64E-09
Cflo_N_g2688	Hypothetical protein	-3.2330	6.17E-05	-3.2280	3.25E-07
Cflo_N_g4443	Trypsin-1	-3.4745	4.23E-05	-3.1988	9.62E-14
Cflo_N_g13662	JHBP, Haemolymph juvenile hormone binding protein (JHBP)	-3.0929	0.00065	-3.0929	5.02E-13
Cflo_N_g5987	Glycerate kinase	-2.5940	0.99998	-3.0548	4.53E-05
Cflo_N_g2277	Acidic mammalian chitinase	-2.9229	0.15386	-2.9391	6.43E-12
Cflo_N_g5844	DUF745 domain containing protein	-2.7241	0.02975	-2.9299	1.39E-06
Cflo_N_g8231	Maltase 1	-2.6227	0.00072	-2.8498	5.58E-11
Cflo_N_g2496	Putative cytochrome P450 6g2/6k1	-2.6571	1.40E-07	-2.7665	2.52E-10
Cflo_N_g10612	Hexamerin-1.1	-4.3845	6.43E-09	-2.7263	3.32E-10
Cflo_N_g2689	Hypothetical protein	-2.7470	0.07220	-2.7048	3.32E-10
Cflo_N_g7901	Hypothetical protein	-2.4791	0.00087	-2.6962	5.07E-10
Cflo_N_g11772	Hypothetical protein	-2.5278	0.99998	-2.6713	7.27E-07

Cflo_N_g6635	Prostatic acid phosphatase	-2.5426	2.02E-05	-2.6655	1.15E-08
Cflo_N_g9682	Putative trypsin-6/Trypsin 3A1	-2.5718	3.19E-05	-2.6007	3.09E-07
Cflo_N_g2331	Putative odorant receptor 13a	-2.2902	0.00885	-2.5625	0.00653
Cflo_N_g11635	Guanine deaminase	-2.5030	0.00072	-2.5611	5.03E-09
Cflo_N_g1918	Phenoloxidase subunit A3	-2.5292	4.71E-05	-2.5504	4.75E-09
Cflo_N_g5236	Protein G12	0.0000	1	-2.4756	1.03E-08
Cflo_N_g13955	Lipase 3	-2.6009	1.07E-05	-2.4736	1.54E-06
Cflo_N_g9926	52 kDa repressor of the inhibitor of the protein kinase	-1.0706	0.99998	-2.4701	7.33E-07
Cflo_N_g12499	Lipase 3	-2.4677	0.00890	-2.4419	2.05E-08
Cflo_N_g6500	Hexamerin-1.1/Hexamerin	0.0000	1	-2.4253	2.10E-08
Cflo_N_g1453	Alpha-glucosidase	-2.3417	0.00065	-2.4149	3.20E-08
Cflo_N_g2547	Trypsin-1-like	-2.0843	0.42093	-2.4105	0.02678
Cflo_N_g8234	Maltase 1	-2.5538	0.03109	-2.4098	4.52E-08
Cflo_N_g10613	Hexamerin	0.0000	1	-2.3145	1.10E-07
Cflo_N_g6021	Larval cuticle protein LCP-17-like	-1.8491	0.68734	-2.3015	0.00029
Cflo_N_g4036	Lysozyme c-1	-1.6453	0.06390	-2.2406	0.00020
Cflo_N_g5292	Organic cation transporter protein	-1.9774	0.00150	-2.2312	1.57E-06
Cflo_N_g1906	Fatty acid synthase	-2.3657	0.05194	-2.2121	0.00015
Cflo_N_g7975	Spermatogenesis-associated protein 17	-2.0869	0.01426	-2.1943	5.26E-05
Cflo_N_g6974	Putative oxidoreductase yrbE	-2.1102	0.15386	-2.1845	7.53E-07
Cflo_N_g8342	probable allantoinase 1-like	-2.1482	0.01267	-2.1694	1.18E-06
Cflo_N_g2305	Sorbitol dehydrogenase	-2.1790	0.56144	-2.1473	1.27E-06
Cflo_N_g12034	Cytochrome P450 4g15/4C1	-2.0261	0.01151	-2.1387	0.00244
Cflo_N_g1543	L-asparaginase	-1.8931	0.02418	-2.1012	0.00043
Cflo_N_g8105	Hypothetical protein	-1.1133	0.79480	-2.0760	0.00150
Cflo_N_g13507	Putative vitellogenin receptor	-1.8283	0.08358	-2.0610	0.02849
Cflo_N_g14172	Probable G-protein coupled receptor Mth-like 1	-2.3634	0.02417	-2.0479	0.00244
Cflo_N_g2490	Alkaline phosphatase 4	-1.9494	0.05053	-2.0398	5.79E-06
Cflo_N_g5968	Solute carrier family 2, facilitated glucose transporter member 8	-2.0571	0.19514	-2.0014	0.02678
Cflo_N_g8960	Lysosomal alpha-mannosidase-like	-1.9328	0.99998	-1.9968	8.88E-06
Cflo_N_g6810	Fatty acyl-CoA reductase 1	-1.9463	0.01151	-1.9598	0.00393
Cflo_N_g3223	Putative fatty acyl-CoA reductase	-2.2873	0.02631	-1.9410	0.72930
Cflo_N_g6775	Hypothetical protein	-1.8274	0.92004	-1.9409	1.94E-05
Cflo_N_g4442	Trypsin-1	-1.8389	0.00615	-1.9250	0.00026
Cflo_N_g13122	Probable G-protein coupled receptor Mth-like 10	-1.7974	0.06391	-1.9189	0.00021
Cflo_N_g3582	5'-nucleotidase	-1.8618	0.99998	-1.8949	4.99E-05
Cflo_N_g8803	Zinc carboxypeptidase A 1	-1.8301	0.65892	-1.8943	3.44E-05
Cflo_N_g2034	Uricase (Fragment)	-1.6250	0.55094	-1.8869	3.93E-05
Cflo_N_g7297	Chymotrypsin-2	-2.6430	0.99998	-1.8860	4.36E-05
Cflo_N_g4555	Peritrophin-1/Peritrophic membrane protein 1 (Fragment)	-1.7974	0.55920	-1.8827	4.03E-05
Cflo_N_g4581	G-protein coupled receptor Mth2	-1.9440	0.00804	-1.8782	0.00036
Cflo_N_g5481	Beta-galactosidase (Fragment)	-1.7721	0.03344	-1.8759	4.99E-05
Cflo_N_g14819	Puromycin-sensitive aminopeptidase N	-1.7912	0.00473	-1.8671	0.00018
Cflo_N_g6857	Flexible cuticle protein 12/Sodium-dependent phosphate transport protein 1	-3.1364	0.99998	-1.8624	5.73E-05

Cflo_N_g4141	Hypothetical protein	-1.7518	0.83597	-1.8591	5.73E-05
Cflo_N_g14282	Peritrophin-1	-1.3993	0.86071	-1.8421	0.00135
Cflo_N_g5542	Cytochrome P450 18a1	-1.8174	0.00083	-1.8360	0.00013
Cflo_N_g832	Cytochrome P450 6k1	-1.9932	0.23318	-1.8336	0.00044
Cflo_N_g42	Hexokinase-2	-1.7519	0.73269	-1.7987	0.00011
Cflo_N_g7278	Trypsin-1/Carboxypeptidase B	-1.3223	0.65100	-1.7741	0.00038
Cflo_N_g7933	G-protein coupled receptor Mth2	-1.7401	0.72428	-1.7736	0.01000
Cflo_N_g6260	Alpha-tocopherol transfer protein	-1.5282	0.24403	-1.7726	0.00197
Cflo_N_g8577	Synaptic vesicle glycoprotein 2B	-1.5834	0.37366	-1.7446	0.00739
Cflo_N_g1853	mucin-6-like	-1.6996	0.18079	-1.7080	0.00034
Cflo_N_g7296	Chymotrypsin-1	1.7200	0.99998	-1.7056	0.00031
Cflo_N_g6298	Zinc metalloprotease zmpB	-1.6305	0.99998	-1.6873	0.00436
Cflo_N_g2306	Lachesin	0.2226	0.99998	-1.6830	0.00040
Cflo_N_g9061	GJ23750/Nucleolar protein 6	0.2153	0.99998	-1.6614	0.00059
Cflo_N_g1762	Probable cytochrome P450 4aa1	-1.9481	0.68734	-1.6546	0.00137
Cflo_N_g13170	Cytochrome P450 6g2/6k1	-1.6154	0.99998	-1.6456	0.00192
Cflo_N_g7543	Sphingomyelin phosphodiesterase	-1.6295	0.55094	-1.6387	0.00067
Cflo_N_g10268	Putative secreted protein	-1.6234	0.00952	-1.6218	0.00088
Cflo_N_g5543	Cytochrome P450 306a1	-1.4667	0.45889	-1.6120	0.04037
Cflo_N_g8893	Probable alpha-ketoglutarate-dependent hypophosphite dioxygenase/MAD2L1-binding protein	-1.6003	0.17584	-1.6112	0.00104
Cflo_N_g2989	Sugar transporter ERD6-like 7	-1.5004	0.14689	-1.5966	0.00228
Cflo_N_g2545	Mite allergen Der p 3/Trypsin-1-like	-2.2138	0.04264	-1.5961	0.59141
Cflo_N_g9686	Brevenin domain containing protein	-1.5071	0.18626	-1.5831	0.00134
Cflo_N_g5053	Membrane alanyl aminopeptidase N	-1.2120	0.99998	-1.5816	0.00582
Cflo_N_g8085	Prostatic acid phosphatase	-1.4688	0.02130	-1.5656	0.00299
Cflo_N_g4435	Dipeptidase 1	-1.4849	0.03128	-1.5538	0.00242
Cflo_N_g3392	Putative gustatory receptor 43a	-1.4521	0.05053	-1.5332	0.03169
Cflo_N_g13852	Hypothetical protein	-1.3278	0.49702	-1.5305	0.04258
Cflo_N_g5954	Bile salt-activated lipase/Esterase E4	-1.4793	0.99998	-1.5269	0.00400
Cflo_N_g833	Cytochrome P450 6k1	-1.4737	0.10167	-1.5060	0.01239
Cflo_N_g4454	Ecdysteroid UDP-glucosyltransferase	-1.4669	0.13879	-1.4991	0.00410
Cflo_N_g9201	Estradiol 17-beta-dehydrogenase 8	-1.5008	0.10690	-1.4989	0.00509
Cflo_N_g13481	Protein takeout	-1.4093	0.65892	-1.4899	0.00364
Cflo_N_g2715	Membrane metallo-endopeptidase-like 1/Endothelin-converting enzyme-like 1	-1.3603	0.12152	-1.4775	0.00430
Cflo_N_g2693	Vitamin K-dependent gamma-carboxylase	-1.2912	0.30850	-1.4696	0.02070
Cflo_N_g5085	Transient receptor potential channel pyrexia	-1.4650	0.10467	-1.4606	0.03301
Cflo_N_g14481	Sugar transporter ERD6-like 8	-1.4258	0.23902	-1.4460	0.00886
Cflo_N_g12976	Adenine phosphoribosyltransferase	-1.3876	0.21638	-1.4437	0.01040
Cflo_N_g13047	RCC1 and BTB domain-containing protein 1	-1.4075	0.24858	-1.4262	0.02292
Cflo_N_g13016	Putative odorant-binding protein A10/Metallophosphoesterase domain-containing protein 1	-0.6728	0.99998	-1.4176	0.00751
Cflo_N_g7215	Retinol dehydrogenase 12	-1.2084	0.73610	-1.4014	0.01350
Cflo_N_g3258	Probable cytochrome P450 9e2/6a13	-1.3350	0.06386	-1.4009	0.00929
Cflo_N_g3056	Peritrophin-1	-1.3578	0.99998	-1.3947	0.00793
Cflo_N_g665	Hypothetical protein	-1.2687	0.66312	-1.3904	0.01061

Cflo_N_g4273	Hypothetical protein	-0.9661	0.99998	-1.3899	0.02635
Cflo_N_g2561	Kynurenine/alpha-aminoadipate aminotransferase mitochondrial	-1.2063	0.26249	-1.3883	0.02292
Cflo_N_g8298	Kv channel-interacting protein 2	-0.0005	0.99998	-1.3780	0.01772
Cflo_N_g12071	Serine-threonine-protein kinase RIO2/Sorbitol dehydrogenase	-1.4610	0.92004	-1.3756	0.00989
Cflo_N_g4365	venom serine protease 34-like/Suppressor of tumorigenicity protein 14	-0.7383	0.99998	-1.3593	0.01325
Cflo_N_g9696	Tenascin-X/von Willebrand factor D and EGF domain-containing protein	-1.4090	0.20056	-1.3593	0.03080
Cflo_N_g4853	L-xylulose reductase	-1.2354	0.15331	-1.3509	0.01373
Cflo_N_g10575	Beta-glucuronidase	-1.2545	0.663312	-1.3477	0.01314
Cflo_N_g5958	Esterase FE4/Para-nitrobenzyl esterase	-1.1851	0.99998	-1.3453	0.02331
Cflo_N_g5956	Para-nitrobenzyl esterase/Liver carboxylesterase 31	-1.0418	0.65121	-1.3279	0.04658
Cflo_N_g793	Calexcitin-2	-1.2982	0.29885	-1.3203	0.01747
Cflo_N_g5456	Septin-4	-1.3333	0.57280	-1.3161	0.01926
Cflo_N_g13757	EcKinase, Ecdysteroid kinase	-1.2153	0.66312	-1.3108	0.01881
Cflo_N_g12949	Activin_recp domain containing protein	-1.1927	0.53534	-1.3038	0.03563
Cflo_N_g5643	Plasma glutamate carboxypeptidase	-1.5278	0.09094	-1.2999	0.02267
Cflo_N_g2156	chymotrypsin 2, Anopheles gambiae-like	-1.1833	0.99998	-1.2890	0.02339
Cflo_N_g8279	Solute carrier family 2, facilitated glucose transporter member 8	-1.2199	0.24858	-1.2766	0.02734
Cflo_N_g9349	Luciferin 4-monooxygenase (Fragment)/4-coumarate--CoA ligase 4	-1.3411	0.41235	-1.2716	0.02899
Cflo_N_g1756	Epididymal secretory protein E1/NPC2 homolog	-1.2807	0.35490	-1.2579	0.02881
Cflo_N_g14330	Cysteine dioxygenase type 1	-1.11	0.27	-1.22	0.051776121
Cflo_N_g14858	Lipase member H-A (Fragment)	-1.21	1.00	-1.20	0.052915274
Cflo_N_g5957	Para-nitrobenzyl esterase	-1.11	1.00	-1.19	0.057722986
Cflo_N_g9796	Very low-density lipoprotein receptor	-1.22	1.00	-1.18	0.056079814
Cflo_N_g10993	Elongation of very long chain fatty acids protein 1	1.6950	0.02462	-0.8571	0.88488
Cflo_N_g4301	Cytochrome b-c1 complex subunit 8	2.1631	0.01212	-0.4229	1
Cflo_N_g11706	Cytochrome P450 6A1	3.0291	2.19E-05	0.2173	1
Cflo_N_g2844	Transposase	2.1694	0.03513	0.2579	1
Cflo_N_g5455	Septin-4	2.5493	0.01458	0.5479	1
Cflo_N_g12492	similar to gag-pol polyprotein	4.0142	0.00473	0.8749	1
Cflo_N_g14504	Lysosomal aspartic protease (Fragment)	1.7223	0.00478	1.0532	0.14249
Cflo_N_g2827	Voltage-dependent calcium channel type A subunit alpha-1	6.5333	0.00340	1.0950	1
Cflo_N_g9088	Chondroitin proteoglycan-2	1.19	0.28	1.18	0.056391144
Cflo_N_g4956	Waprin-Phi1	0.22	1.00	1.19	0.057255246
Cflo_N_g11036	ATP-dependent RNA helicase	0.27	1.00	1.20	0.050769243
Cflo_N_g925	Ornithine decarboxylase/Vacuolar protein sorting-associated protein 37B	1.3388	0.99998	1.2021	0.04507
Cflo_N_g2002	Tetraspanin-9	1.2918	0.99998	1.2072	0.04316
Cflo_N_g12186	Silk fibroin	1.3765	0.99998	1.2370	0.03298
Cflo_N_g12187	Silk fibroin	1.3505	0.99998	1.2431	0.03136
Cflo_N_g12144	ATP-binding cassette sub-family A member 13	1.3234	0.99998	1.2707	0.02555
Cflo_N_g977	L-lactate dehydrogenase	1.2541	0.28466	1.2710	0.04658
Cflo_N_g14413	NF-κB inhibitor cactus	1.3138	0.99998	1.2841	0.02291
Cflo_N_g9521	Transient receptor potential channel pyrexia	1.2841	0.16231	1.2979	0.03420

Cflo_N_g12523	15-hydroxyprostaglandin dehydrogenase [NAD+]	1.9992	0.63184	1.3040	0.01971
Cflo_N_g9101	Kynurenine/alpha-aminoadipate aminotransferase mitochondrial	1.4095	0.32539	1.3147	0.01772
Cflo_N_g6458	Leucine-rich repeat-containing protein 20	1.3020	0.61519	1.3290	0.01595
Cflo_N_g3367	Protein msta, isoform A	1.52	1.00	1.33	0.054702949
Cflo_N_g6082	NF-κB p110 subunit (Relish)	1.5379	0.99998	1.3406	0.01510
Cflo_N_g5763	Tyrocidine synthetase 3	1.4563	0.11693	1.3872	0.00957
Cflo_N_g15550	Maltase 1	1.4315	0.04053	1.3936	0.01657
Cflo_N_g6700	Hypothetical protein	1.5190	0.99998	1.3958	0.00783
Cflo_N_g455	Apoptosis-inducing factor 3	1.5133	0.10546	1.3985	0.00874
Cflo_N_g13089	Serpin B10	1.0089	0.97948	1.4033	0.01747
Cflo_N_g13654	5-aminolevulinate synthase, erythroid-specific, mitochondrial	1.4768	0.98136	1.4083	0.00837
Cflo_N_g5176	Retrotransposable element (Fragment)	1.4615	0.99998	1.4148	0.04829
Cflo_N_g863	MIT domain-containing protein 1	1.5071	0.99998	1.4222	0.00896
Cflo_N_g3993	Brain-specific angiogenesis inhibitor 1/Hypothetical protein	2.3063	0.99998	1.4346	0.00609
Cflo_N_g5162	Hypothetical protein	1.6209	0.02591	1.4355	0.01657
Cflo_N_g6275	Hypothetical protein	1.4803	0.27379	1.4421	0.00582
Cflo_N_g8790	Hypothetical protein	1.6877	0.88437	1.4457	0.00497
Cflo_N_g115	Innexin shaking-B/Hypothetical protein	1.5382	0.06252	1.4700	0.01571
Cflo_N_g4449	Hypothetical protein	1.5469	0.12494	1.4771	0.00444
Cflo_N_g9207	Lysosomal-trafficking regulator	1.5782	0.17653	1.4865	0.00369
Cflo_N_g9833	Hypothetical protein	1.5487	0.01655	1.4890	0.00402
Cflo_N_g11034	Lysosomal acid phosphatase/ testicular acid phosphatase homolog	1.8278	0.99998	1.4898	0.00385
Cflo_N_g8897	Haemolymph lipopolysaccharide-binding protein/Tyrosine-protein kinase Btk29A	0.7061	0.99998	1.5043	0.00370
Cflo_N_g11948	Acyl-CoA desaturase	1.5988	0.99998	1.5568	0.03298
Cflo_N_g13323	AAEL006341-PA	1.6675	0.99998	1.5803	0.00422
Cflo_N_g8787	DUF745 domain containing protein	1.7736	0.01566	1.5855	0.00508
Cflo_N_g8989	Serine protease snake/Dynein beta chain, ciliary	1.3468	0.99998	1.5875	0.00129
Cflo_N_g7705	Cell division cycle protein 16-like protein/Cytoplasmic polyadenylation element-binding protein 1-B	1.4224	0.36289	1.6149	0.00183
Cflo_N_g11332	Hormone-sensitive lipase	1.6945	0.02002	1.6622	0.00085
Cflo_N_g6505	Succinate-semialdehyde dehydrogenase, mitochondrial	1.9471	0.02524	1.7016	0.00036
Cflo_N_g14403	Serine proteinase stubble	1.8184	0.43716	1.7508	0.00018
Cflo_N_g4958	COG2214 DnaJ-class molecular chaperone	1.8272	0.01162	1.7798	0.00015
Cflo_N_g4389	Putative serine/threonine phosphatase	1.3169	0.99998	1.8022	0.00035
Cflo_N_g11917	Lipoma-preferred partner-like protein	2.4033	0.99998	1.8550	6.48E-05
Cflo_N_g8262	Vitellogenin	1.9387	0.08358	1.8634	0.00561
Cflo_N_g7691	Serine proteinase stubble	1.8802	0.48667	1.8753	4.32E-05
Cflo_N_g11683	WD repeat-containing protein 68	0.4319	0.99998	1.8911	5.10E-05
Cflo_N_g11919	Lysosomal acid phosphatase	0.4722	0.99998	1.9016	3.89E-05
Cflo_N_g8309	Exonuclease GOR (Fragment)	2.3778	0.00050	1.9417	3.08E-05
Cflo_N_g7272	Vitellogenic carboxypeptidase	1.9805	0.00051	1.9441	3.93E-05
Cflo_N_g3376	Calcium-independent phospholipase A2-gamma	1.9483	0.00279	1.9521	2.41E-05
Cflo_N_g9064	Sodium-independent sulfate anion transporter	2.2177	0.22140	1.9578	4.53E-05

Cflo_N_g4772	bone morphogenetic protein 2-B-like	2.1072	0.99998	1.9649	0.00128
Cflo_N_g4603	ETS-like proteinous factor	1.9753	0.00509	1.9687	2.00E-05
Cflo_N_g7277	Cationic trypsin-3/Lectizyme	2.0862	0.92464	1.9709	1.22E-05
Cflo_N_g2995	Cytochrome P450 307a1	1.8890	0.26170	1.9957	0.00017
Cflo_N_g7775	Endocuticle structural glycoprotein SgAbd-1/Pupal cuticle protein Edg-78E	2.4022	0.03344	2.0224	0.52277
Cflo_N_g13783	Apolipoprotein D/LIM domain kinase 1	2.3093	0.99998	2.0279	5.79E-06
Cflo_N_g11850	Three prime repair exonuclease 2	1.9360	0.04391	2.0584	4.94E-06
Cflo_N_g501	Transposable element Tc3 transposase	4.1169	3.30E-06	2.0813	1
Cflo_N_g4516	Kazal-type serine protease inhibitor	2.2226	0.39125	2.0857	2.96E-06
Cflo_N_g14238	Major royal jelly protein 1	2.1969	0.00170	2.0872	3.81E-06
Cflo_N_g7438	Coagulation factor IX (Fragment)/Limulus clotting factor C	2.1797	0.00545	2.1389	1.57E-06
Cflo_N_g1062	Hypothetical protein	2.3655	0.00279	2.1544	1.57E-06
Cflo_N_g4257	Leukocyte elastase inhibitor	2.1938	0.02861	2.1593	1.14E-06
Cflo_N_g8403	Protein Dom3Z	2.1469	0.00016	2.1791	5.56E-06
Cflo_N_g3960	Vanin-like protein 1/Cytochrome b-c1 complex subunit Rieske, mitochondrial	2.2729	0.01151	2.2295	4.69E-07
Cflo_N_g7396	Hypothetical protein	2.3506	0.99998	2.2437	3.92E-06
Cflo_N_g2024	L-ascorbate oxidase/Laccase-4	2.4754	1.90E-05	2.2441	1.44E-06
Cflo_N_g4388	putative leucine-rich repeat-containing protein DDB_G0290503-like	2.3852	0.00055	2.3769	1.98E-07
Cflo_N_g7714	Transferrin	2.4811	0.99998	2.4005	3.06E-08
Cflo_N_g10198	Translation initiation factor IF-2	2.5359	0.00012	2.4120	2.92E-08
Cflo_N_g839	Putative methyltransferase METT10D	2.5791	0.02434	2.4618	1.59E-08
Cflo_N_g13573	Protein SERAC1	2.9268	7.48E-07	2.4907	1.59E-08
Cflo_N_g8442	Serine protease persephone/snake	2.6004	0.00163	2.4970	8.47E-09
Cflo_N_g11682	Transmembrane protein 205	2.8434	0.00340	2.5052	1.03E-08
Cflo_N_g11918	Protein Malvolio	2.7600	0.28466	2.5520	3.90E-09
Cflo_N_g344	Hypothetical protein	2.8922	1.17E-06	2.5627	3.38E-06
Cflo_N_g5742	Beta-1,3-glucan-binding protein	2.7400	0.00473	2.6715	6.03E-10
Cflo_N_g103	Peptidoglycan-recognition protein-LB	2.6626	6.97E-08	2.6959	6.03E-10
Cflo_N_g3614	Transient receptor potential channel pyrexia	2.3813	0.00371	2.7142	4.83E-08
Cflo_N_g13479	Circadian clock-controlled protein	2.5709	0.00135	2.7252	2.43E-07
Cflo_N_g3954	Fatty acyl-CoA reductase 1	4.1506	0.00066	2.7334	0.32525
Cflo_N_g5252	Protein G12	2.9096	0.99998	2.7712	1.03E-10
Cflo_N_g7698	Hypothetical protein	3.0306	0.99998	2.8327	1.48E-09
Cflo_N_g3252	Serine protease inhibitor dipetalogastin	3.5301	8.71E-05	2.9051	0.00128
Cflo_N_g7790	Hypothetical protein	3.1300	5.27E-05	2.9320	6.38E-07
Cflo_N_g6072	PREDICTED: titin-like	3.2056	2.22E-06	2.9771	2.13E-05
Cflo_N_g8188	Transposable element Tcb1 transposase (Fragment)	5.0757	5.17E-07	3.0073	0.57344
Cflo_N_g4390	probable salivary secreted peptide-like	3.0460	0.99998	3.0497	9.06E-13
Cflo_N_g8441	Serine protease snake 1	3.09	1.00	3.08	0.050419592
Cflo_N_g5222	Tyrosine 3-monooxygenase	3.1989	0.31048	3.1181	2.92E-13
Cflo_N_g8315	G-protein coupled receptor Mth2	2.9568	0.00070	3.1703	0.00569
Cflo_N_g790	Synaptic vesicle glycoprotein 2B	3.2237	3.87E-08	3.1768	3.45E-13
Cflo_N_g769	Synaptic vesicle glycoprotein 2B	3.3794	1.67E-09	3.2855	8.41E-11
Cflo_N_g5800	Cytochrome b5	3.7275	6.57E-13	3.6695	1.99E-16

Cflo_N_g6985	Parathyroid hormone/parathyroid hormone-related peptide receptor	3.7699	1.27E-10	3.6754	2.69E-17
Cflo_N_g11755	Aromatic-L-amino-acid decarboxylase	3.6727	6.72E-12	3.6791	3.29E-17
Cflo_N_g14774	HIV Tat-specific factor 1-like protein	4.3359	0.03558	3.9927	3.76E-20
Cflo_N_g8312	Defensin-2	4.7614	2.77E-08	4.5942	1.12E-25
Cflo_N_g8596	Esterase FE4	4.7663	0.01973	4.7211	8.25E-27
Cflo_N_g7715	WD repeat-containing protein C10orf79	4.7947	0	5.0311	4.21E-17
Cflo_N_g12631	Serine proteinase stubble	5.8259	0.00000	5.1857	1.26E-15
Cflo_N_g531	Aminopeptidase N	4.0407	0.00551	5.2188	0.05833
Cflo_N_g6748	Hypothetical protein	8.1100	0	7.7367	6.48E-38
Cflo_N_g10990	EcKinase, Ecdysteroid kinase	2.7340	0.01031	no data available	no data available
Cflo_N_g8411	Transposase	4.3521	9.38E-05	no data available	no data available
Cflo_N_g8963	Structural maintenance of chromosomes protein 1A	5.4285	0.04053	no data available	no data available
-	Defensin prepropeptide	-1.1120	0.99239	-1.5529	0.00320

Appendix VII. List of *C. floridanus* specific hypothetical genes expressed differentially after immune challenge including KOG annotations.

Re-annotation ID	DESeq log2(fc)	Signal peptide	Domain/motif	Transmembrane region	KOG Category	KOG annotation	Class description
Cflo_N_g2688	-3.22796	No	Six Proline rich extensin like signature motifs (Prints:PR01217)	No	KOG3514	Neurexin III-alpha	Signal transduction mechanisms
Cflo_N_g2689	-2.70482	Yes (SignalP)	No	No	KOG4701	Chitinase	Cell wall/membrane/envelope biogenesis
Cflo_N_g4141	-1.85908	No	No	No	ND	-	-
Cflo_N_g6775	-1.94089	No	No	Yes	ND	-	-
Cflo_N_g3392	-1.53322	No	7tm_7 Chemosensory receptor	Yes	ND	-	-
Cflo_N_g11772	-2.67132	No	No	No	ND	-	-
Cflo_N_g8235	-4.59815	No	No	No	ND	-	-
Cflo_N_g5992	-5.20638	Yes (SignalP)	Membrane lipoprotein lipid attachment site, localization signal (Prosite:PS51257)	No	ND	-	-
Cflo_N_g9926	-2.47007	No	THAP (C2CH-type zinc finger) DNA-binding domain	No	ND	-	-

Note – ND – not determined.

Appendix VIII. Genes differentially expressed (DEGs) after immune challenge. The table presents a comparison of expression data of 37 selected immune genes based on the analysis of the Illumina sequencing data (Cuffdiff and DESeq) and the corresponding qRT-PCR data (table prepared by cooperation partners). Significance of resulting log2(x-fold change) values is given by asterisks (* *p*-value/*q*-value ≤0.05; ** ≤0.01; *** ≤0.001; n.s. for non-significant results).

		qRT data					RNAseq			
		Larvae L2		Worker W2		L2 and W2				
Re-annotation ID	Gene Symbol	log2(x-fold change L2)	significance	log2(x-fold change W2)	significance	log2(x-fold change L2+W2)	Cuffdiff log2(x-fold change)	significance	DESeq log2(x-fold change)	significance
	Housekeeping Gene									
Cflo_N_g6999	rpL32	0.0841	n.s.	0.0144	n.s.	0.0426	-0.0564	n.s.	-0.1821	n.s.
	Potential Immune Genes									
Cflo_N_g9950	scav	-0.5353	n.s.	-1.9434	**	-1.0893	-2.7179	n.s.	-1.1126	n.s.
Cflo_N_g1918	Posub	-1.4739	**	-0.9434	n.s.	-1.1844	-2.5292	***	-2.5504	***
Cflo_N_g9172	CathD	0.6041	n.s.	0.2388	n.s.	0.4330	-4.2058	n.s.	inf	n.s.
Cflo_N_g8803	zcp	-0.7131	*	0.9260	n.s.	0.3219	-1.8089	n.s.	-1.8256	***
Cflo_N_g2215	hpchit	-2.9434	*	1.0000	**	0.0976	-8.0072	***	-7.9683	***
Cflo_N_g5542	cP45018a1	-0.3585	n.s.	0.6599	n.s.	0.2388	-1.8174	***	-1.8360	***
Cflo_N_g10836	sushi	-0.1203	n.s.	-0.1520	n.s.	-0.1361	-1.0499	n.s.	-1.1145	n.s.
Cflo_N_g907	chymo	-2.3219	*	0.3561	n.s.	-0.4344	-6.1752	***	-6.2845	***
Cflo_N_g14858	lip	-1.0291	*	-0.0291	n.s.	-0.4540	-1.2117	n.s.	-1.1990	n.s.
Cflo_N_g2277	chito	-1.2863	*	0.0144	n.s.	-0.4941	-2.9229	n.s.	-2.9391	***
Cflo_N_g6500	hex	-2.0000	*	0.0704	n.s.	-0.6215	0.0000	n.s.	-2.4253	***
Cflo_N_g1319	hppaci	-2.0000	*	-0.1361	n.s.	-0.7859	-6.9677	***	-7.4057	***
Cflo_N_g622	MPI	2.5656	*	-0.1203	n.s.	1.7740	1.0954	n.s.	0.9240	n.s.
Cflo_N_g14239	yellow	1.1763	***	1.2079	**	1.1890	1.1576	n.s.	1.1196	n.s.
Cflo_N_g4920	SOCS2	2.2510	**	0.7485	*	1.6871	1.2949	n.s.	1.0937	n.s.
Cflo_N_g14413	cact1	2.8777	n.s.	1.5705	**	2.3674	1.3138	n.s.	1.2841	*
Cflo_N_g8262	vitel	0.1506	n.s.	3.1985	n.s.	2.3646	1.9387	n.s.	1.8634	**
Cflo_N_g7714	transf	3.5969	***	0.5656	n.s.	2.7634	2.4811	n.s.	2.4005	***
Cflo_N_g9484	hp70940	2.4383	*	3.0531	***	2.7782	0.4320	n.s.	0.1401	n.s.
Cflo_N_g6985	PHR	5.1408	***	2.7592	**	4.3944	3.7699	***	3.6754	***
Cflo_N_g8597	ester	5.3409	***	0.8875	n.s.	4.4053	4.7663	*	4.7211	n.s.
Cflo_N_g6748	hp67112	9.8711	***	3.2388	n.s.	8.8855	8.1100	***	7.7367	***

Cflo_N_g103	pgrp-LB	3.5934	***	1.6508	***	2.9260	2.6626	***	2.6959	***
Cflo_N_g8526	pgrp-2	2.7949	***	0.9486	**	2.1506	0.0000	n.s.	0.9069	n.s.
Cflo_N_g6082	rel	1.3618	*	2.1731	***	1.8237	1.5379	n.s.	1.3406	*
Cflo_N_g14777	hym	8.6782	***	4.1538	**	7.7396	no data	no data	no data	no data
Cflo_N_g8312	def	3.3743	***	2.8758	**	3.1473	4.7614	***	4.5942	***
Cflo_N_g5222	tyrOH	3.5160	**	2.9165	***	3.2464	3.1989	n.s.	3.1181	***
Cflo_N_g7345	tep1	4.4995	***	-0.2515	n.s.	3.5521	0.4746	n.s.	0.3879	n.s.
Cflo_N_g197	nos1	1.3618	*	-0.1047	n.s.	0.8074	0.5499	n.s.	0.4668	n.s.
Cflo_N_g1023	lyso i-type	0.2388	n.s.	1.1440	**	0.7655	1.2310	n.s.	1.1129	n.s.
Cflo_N_g7081	imdK1	1.0566	**	-0.1047	n.s.	0.5945	0.1814	n.s.	0.0472	n.s.
Cflo_N_g3305	dorsal	0.8480	*	-0.0439	n.s.	0.4647	0.1960	n.s.	0.1417	n.s.
Cflo_N_g14504	lap61281	0.3896	n.s.	0.2987	n.s.	0.3448	1.5891	**	1.5421	n.s.
Cflo_N_g9451	mapkkk7	0.5656	*	-0.0145	n.s.	0.2987	0.4391	n.s.	0.2613	n.s.
Cflo_N_g4492	tep2	0.3785	n.s.	0.1375	n.s.	0.2630	0.4650	n.s.	0.4110	n.s.
Cflo_N_g9745	tep3	-0.0291	n.s.	0.1506	n.s.	0.0566	0.2594	n.s.	0.2015	n.s.

Appendix IX. Differentially expressed hubs and bottlenecks in top20 % of proteins in *C. floridanus* signalome.

Re-annotation ID	Annotation	Hub/Bottleneck
Cflo_N_g6260	neurogenic locus Notch protein-like	Hub+Bottleneck
Cflo_N_g9349	Luciferin 4-monooxygenase	Hub
Cflo_N_g14413	NF-κB inhibitor cactus	Hub+Bottleneck
Cflo_N_g6082	NF-κB p110 subunit	Hub+Bottleneck
Cflo_N_g9686	Brevenin domain containing protein	Hub
Cflo_N_g5742	Beta-1,3-glucan-binding protein (GNBP)	Bottleneck

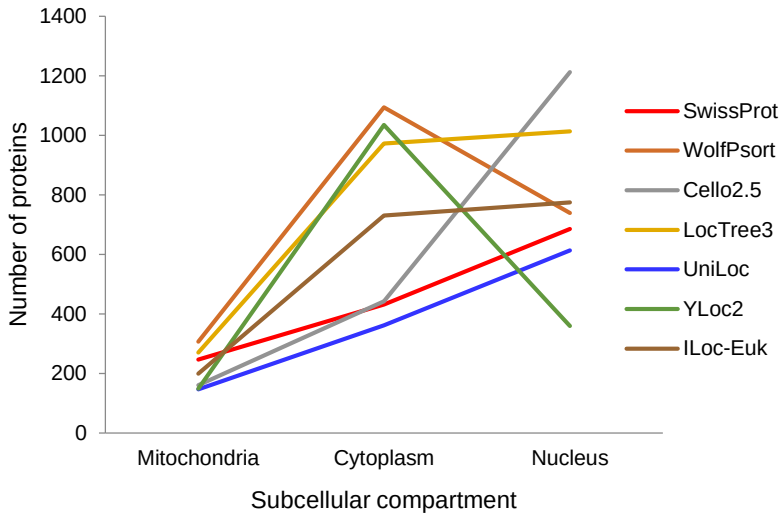
Appendix X. Differentially expressed hubs and bottlenecks in top 20 % of proteins in IISC network of *C. floridanus*.

Re-annotation ID	Annotation	Hub/Bottleneck
Cflo_N_g6260	neurogenic locus Notch protein-like	Hub+Bottleneck
Cflo_N_g6082	NF-κB p110 subunit	Hub+Bottleneck
Cflo_N_g7775	Endocuticle structural glycoprotein SgAbd-1	Hub
Cflo_N_g977	L-lactate dehydrogenase	Hub+Bottleneck
Cflo_N_g9349	Luciferin 4-monooxygenase	Hub
Cflo_N_g11706	Cytochrome P450 6A1	Hub+Bottleneck
Cflo_N_g14413	NF-κB inhibitor cactus	Hub
Cflo_N_g5222	Tyrosine 3-monooxygenase	Hub+Bottleneck
Cflo_N_g10990	EcKinase, Ecdysteroid kinase	Hub+Bottleneck
Cflo_N_g2277	Acidic mammalian chitinase	Hub+Bottleneck
Cflo_N_g6859	Flexible cuticle protein 12	Hub
Cflo_N_g8441	Serine protease snake 1	Hub+Bottleneck
Cflo_N_g11036	ATP-dependent RNA helicase	Hub
Cflo_N_g9686	Brevenin domain containing protein	Hub
Cflo_N_g5742	Beta-1,3-glucan-binding protein (GNBP)	Hub+Bottleneck
Cflo_N_g13089	Serpin B10	Hub+Bottleneck
Cflo_N_g4555	Peritrophic membrane protein 1	Hub+Bottleneck
Cflo_N_g455	Apoptosis-inducing factor 3	Hub+Bottleneck

Appendix XI. Annotation and expression value of the proteins involved in conserved active modules.

Re-annotation ID	Annotation	DESeq log2FC	Larvae haemolymph log2LFQ intensity
Cflo_N_g8541	3-phosphoinositide-dependent protein kinase 1		
Cflo_N_g1453	Alpha-glucosidase	-2.4149	-2.75
Cflo_N_g1198	E3 ubiquitin-protein ligase NRDP1		
Cflo_N_g9576	FK506-binding protein 4		
Cflo_N_g293	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing] 2		-2.02
Cflo_N_g7863	Heat shock protein 90		
Cflo_N_g1023	Lysozyme		
Cflo_N_g4036	Lysozyme c-1	-2.2406	
Cflo_N_g8234	Maltase 1	-2.4098	-3.86
Cflo_N_g15545	Maltase 2		-2.38
Cflo_N_g13662	Haemolymph juvenile hormone binding protein (JHBP)	-3.0929	-3.8
Cflo_N_g10587	Serine/threonine-protein kinase polo		
Cflo_N_g2544	Trypsin-1	-3.5981	-2.17
Cflo_N_g9772	DNA-directed RNA polymerases I and III subunit RPAC1		

Appendix XII. Number of single compartment localized proteins of *C. floridanus* (as suggested by Swiss-Prot orthology for three compartments) present in the interactome and comparisons of localization prediction for these proteins with six localization prediction tools.



Bibliography

Aebersold, R., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422, 198-207.

Agaisse, H., Petersen, U.-M., Boutros, M., Mathey-Prevot, B., Perrimon, N., 2003. Signaling role of hemocytes in *Drosophila* JAK/STAT-dependent response to septic injury. Dev. Cell 5, 441-450.

Aguilar, R., Jedlicka, A.E., Mintz, M., Mahairaki, V., Scott, A.L., Dimopoulos, G., 2005. Global gene expression analysis of *Anopheles gambiae* responses to microbial challenge. Insect Biochem. Mol. Biol. 35, 709-719.

Albinsson, B., Kidd, A.H., 1999. Adenovirus type 41 lacks an RGD α v-integrin binding motif on the penton base and undergoes delayed uptake in A549 cells. Virus Res. 64, 125-136.

Altincicek, B., Vilcinskas, A., 2006. Metamorphosis and collagen-IV-fragments stimulate innate immune response in the greater wax moth, *Galleria mellonella*. Dev. Comp. Immunol. 30, 1108-1118.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol. 11, R106.

Anselme, C., Perez-Brocal, V., Vallier, A., Vincent-Monegat, C., Charif, D., Latorre, A., Moya, A., Heddi, A., 2008. Identification of the weevil immune genes and their expression in the bacteriome tissue. BMC Biol. 6, 43.

Aronstein, K.A., Murray, K.D., Saldivar, E., 2010. Transcriptional responses in honey bee larvae infected with chalkbrood fungus. BMC Genomics 11, 1.

Ashok, Y., 2009. Drosophila toll pathway: the new model. Sci. Signal. 2, jc1-jc1.

Ayres, J.S., Schneider, D.S., 2009. The role of anorexia in resistance and tolerance to infections in Drosophila. PLoS Biol. 7, e1000150.

Bader, G.D., Betel, D., Hogue, C.W., 2003. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31, 248-250.

Baker, M., 2012. De novo genome assembly: what every biologist should know. Nat. Methods 9, 333-337.

Baldwin, M.A., 2004. Protein identification by mass spectrometry: issues to be considered. Mol. Cell. Proteomics 3, 1-9.

Batada, N.N., Hurst, L.D., Tyers, M., 2006. Evolutionary and physiological importance of hub proteins. PLoS Comp. Biol. 2, e88.

Bauer-Mehren, A., Furlong, L.I., Sanz, F., 2009. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol. Syst. Biol. 5.

Beckage, N.E., 2011. Insect immunology. Academic press.

Belda, E., Moya, A., Silva, F.J., 2005. Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. Mol. Biol. Evol. 22, 1456-1467.

Bergeret, E., Perrin, J., Williams, M., Grunwald, D., Engel, E., Thevenon, D., Taillebourg, E., Bruckert, F., Cosson, P., Fauvarque, M.-O., 2008. TM9SF4 is required for Drosophila cellular immunity via cell adhesion and phagocytosis. J. Cell. Sci. 121, 3325-3334.

Bhavsar, A.P., Guttman, J.A., Finlay, B.B., 2007. Manipulation of host-cell pathways by bacterial pathogens. Nature 449, 827-834.

Bilder, D., Perrimon, N., 2000. Localization of apical epithelial determinants by the basolateral PDZ protein Scribble. Nature 403, 676-680.

Bischoff, V., Vignal, C., Duvic, B., Boneca, I.G., Hoffmann, J.A., Royet, J., 2006. Downregulation of the Drosophila immune response by peptidoglycan-recognition proteins SC1 and SC2. PLoS Pathog. 2, e14.

Blandin, S., Levashina, E.A., 2004. Thioester-containing proteins and insect immunity. Mol. Immunol. 40, 903-908.

Boman, H.G., Hultmark, D., 1987. Cell-free immunity in insects. Annu. Rev. Microbiol. 41, 103-126.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., et al., 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329, 1068-1071.

Bou Aoun, R., Hetru, C., Troxler, L., Doucet, D., Ferrandon, D., Matt, N., 2010. Analysis of thioester-containing proteins during the innate immune response of *Drosophila melanogaster*. J. Innate Immun. 3, 52-64.

Bouzid, W., Verdenaud, M., Klopp, C., Ducancel, F., Noirot, C., Vétillard, A., 2014. De Novo sequencing and transcriptome analysis for *Tetramorium bicarinatum*: a comprehensive venom gland transcriptome analysis from an ant species. BMC Genomics 15, 1.

Brey, P.T., Lee, W.-J., Yamakawa, M., Koizumi, Y., Perrot, S., Francois, M., Ashida, M., 1993. Role of the integument in insect immunity: epicuticular abrasion and induction of cecropin synthesis in cuticular epithelial cells. Proc. Natl. Acad. Sci. USA 90, 6275-6279.

Bryant, D.W., Jr., Priest, H.D., Mockler, T.C., 2012. Detection and quantification of alternative splicing variants using RNA-seq. Methods Mol. Biol. 883, 97-110.

Buensuceso, C.S., Woodside, D., Huff, J.L., Plopper, G.E., O'Toole, T.E., 2001. The WD protein Rack1 mediates protein kinase C and integrin-dependent cell migration. J. Cell. Sci. 114, 1691-1698.

Bulet, P., Stocklin, R., Menin, L., 2004. Anti-microbial peptides: from invertebrates to vertebrates. Immunol. Rev. 198, 169-184.

Burridge, K., Wennerberg, K., 2004. Rho and Rac take center stage. Cell 116, 167-179.

Carbonell, G.V., Falcón, R., Yamada, A.T., da Fonseca, B.A.L., Yano, T., 2004. Morphological and intracellular alterations induced by *Serratia marcescens* cytotoxin. Res. Microbiol. 155, 25-30.

Cerenius, L., Söderhäll, K., 2004. The prophenoloxidase-activating system in invertebrates. Immunol. Rev. 198, 116-126.

Chelvanayagam, G., Parker, M.W., Board, P., 2001. Fly fishing for GSTs: a unified nomenclature for mammalian and insect glutathione transferases. Chem. Biol. Interact. 133, 256-260.

Choi, D.-S., Young, H., McMahon, T., Wang, D., Messing, R.O., 2003. The mouse RACK1 gene is regulated by nuclear factor-κB and contributes to cell survival. Mol. Pharmacol. 64, 1541-1548.

Choi, H.K., Choi, K.H., Kramer, K.J., Muthukrishnan, S., 1997. Isolation and characterization of a genomic clone for the gene of an insect molting enzyme, chitinase. Insect Biochem. Mol. Biol. 27, 37-47.

Chu, Y., Corey, D.R., 2012. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 22, 271-274.

Chu, Y., Liu, Y., Shen, D., Hong, F., Wang, G., An, C., 2015. Serine proteases SP1 and SP13 mediate the melanization response of Asian corn borer, *Ostrinia furnacalis*, against entomopathogenic fungus *Beauveria bassiana*. J. Invertebr. Pathol. 128, 64-72.

Colgan, T.J., Carolan, J.C., Bridgett, S.J., Sumner, S., Blaxter, M.L., Brown, M.J., 2011. Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, *Bombus terrestris*. BMC Genomics 12, 623.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation,

visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676.

Cooper, M.D., Alder, M.N., 2006. The evolution of adaptive immune systems. Cell 124, 815-822.

Costa, V., Angelini, C., De Feis, I., Ciccodicola, A., 2010. Uncovering the complexity of transcriptomes with RNA-Seq. J. Biomed. Biotechnol. 2010, 853916.

Cottrell, J.S., 2011. Protein identification using MS/MS data. Journal of proteomics 74, 1842-1851.

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367-1372.

Cremer, S., Armitage, S.A., Schmid-Hempel, P., 2007. Social immunity. Current Biol. 17, R693-R702.

Cronin, S.J., Nehme, N.T., Limmer, S., Liegeois, S., Pospisilik, J.A., Schramek, D., Leibbrandt, A., de Matos Simoes, R., Gruber, S., Puc, U., 2009. Genome-wide RNAi screen identifies genes involved in intestinal pathogenic bacterial infection. Science 325, 340-343.

Daley, W.P., Yamada, K.M., 2013. ECM-modulated cellular dynamics as a driving force for tissue morphogenesis. Curr. Opin. Genet. Dev. 23, 408-414.

Davies, S.-A., Overend, G., Sebastian, S., Cundall, M., Cabrero, P., Dow, J.A., Terhzaz, S., 2012. Immune and stress response 'cross-talk'in the *Drosophila* Malpighian tubule. J. Insect Physiol. 58, 488-497.

De la Vega, H., Specht, C., Liu, Y., Robbins, P., 1998. Chitinases are a multi-gene family in *Aedes*, *Anopheles* and *Drosophila*. Insect Mol. Biol.7, 233-239.

Devos, D., Valencia, A., 2001. Intrinsic errors in genome annotation. Trends Genet. 17, 429-431.

Dinglasan, R.R., Devenport, M., Florens, L., Johnson, J.R., McHugh, C.A., Donnelly-Doman, M., Carucci, D.J., Yates, J.R., 3rd, Jacobs-Lorena, M., 2009. The *Anopheles gambiae* adult midgut peritrophic matrix proteome. Insect Biochem. Mol. Biol. 39, 125-134.

Douglas, A.E., Bouvaine, S., Russell, R.R., 2011. How the insect immune system interacts with an obligate symbiotic bacterium. Proc. R. Soc. Lond., B, Biol. Sci. 278, 333-338.

Dow, L., Elsum, I., King, C., Kinross, K., Richardson, H., Humbert, P., 2008. Loss of human Scribble cooperates with H-Ras to promote cell invasion through deregulation of MAPK signalling. Oncogene 27, 5988-6001.

Du, Z.-Q., Ren, Q., Zhao, X.-F., Wang, J.-X., 2009. A double WAP domain (DWD)-containing protein with proteinase inhibitory activity in Chinese white shrimp, *Fenneropenaeus chinensis*. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 154, 203-210.

Dyer, M.D., Neff, C., Dufford, M., Rivera, C.G., Shattuck, D., Bassaganya-Riera, J., Murali, T., Sobral, B.W., 2010. The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. PloS one 5, e12089.

Eleftherianos, I., More, K., Spivack, S., Paulin, E., Khojandi, A., Shukla, S., 2014. Nitric Oxide Levels regulate the immune response of *Drosophila melanogaster* reference laboratory strains to bacterial infections. Infect. Immun. 82, 4169-4181.

Eleftherianos, I., Revenis, C., 2011. Role and importance of phenoloxidase in insect hemostasis. J. Innate Immun. 3, 28-33.

Elsbach, P., 1994. Bactericidal permeability-increasing protein in host defence against gram-negative bacteria and endotoxin, Antimicrobial Peptides, Ciba Foundation Symposium, pp. 176-187.

Elsbach, P., Weiss, J., Levy, O., 1994. Integration of antimicrobial host defenses: role of the bactericidal/permeability-increasing protein. Trends Microbiol. 2, 324-328.

Elsik, C.G., Worley, K.C., Bennett, A.K., Beye, M., Camara, F., Childers, C.P., de Graaf, D.C., Debyser, G., Deng, J., Devreese, B., et al., Honey Bee Genome Sequencing, C., 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15, 86.

Engel, P., Moran, N.A., 2013. The gut microbiota of insects - diversity in structure and function. FEMS Microbiol. Rev. 37, 699-735.

Esterházy, D., King, M.S., Yakovlev, G., Hirst, J., 2008. Production of reactive oxygen species by complex I (NADH: Ubiquinone Oxidoreductase) from *Escherichia coli* and comparison to the enzyme from mitochondria. Biochemistry 47, 3964-3971.

Evans, J., Aronstein, K., Chen, Y., Hetru, C., Imler, J.L., Jiang, H., Kanost, M., Thompson, G., Zou, Z., Hultmark, D., 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. Insect Mol. Biol.15, 645-656.

Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I.J., 2013. SignaLink 2–a signaling pathway resource with multi-layered regulatory networks. BMC Syst. Biol. 7, 1.

Feder, M., Walser, J.C., 2005. The biological limitations of transcriptomics in elucidating stress and stress responses. J. Evol. Biol. 18, 901-910.

Fellous, S., Lazzaro, B.P., 2011. Potential for evolutionary coupling and decoupling of larval and adult immune gene expression. Mol. Ecol. 20, 1558-1567.

Ferrandon, D., Imler, J.L., Hetru, C., Hoffmann, J.A., 2007. The *Drosophila* systemic immune response: sensing and signalling during bacterial and fungal infections. Nat. Rev. Immunol. 7, 862-874.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., 2013. Pfam: the protein families database. Nucleic Acids Res., gkt1223.

Fjell, C.D., Hancock, R.E., Cherkasov, A., 2007. AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 23, 1148-1155.

Foley, E., O'Farrell, P.H., 2003. Nitric oxide contributes to induction of innate immune responses to gram-negative bacteria in *Drosophila*. Genes Dev. 17, 115-125.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., 2005. Protein interaction mapping: a *Drosophila* case study. Genome Res. 15, 376-384.

Fre, S., Bardin, A., Robine, S., Louvard, D., 2011. Notch signaling in intestinal homeostasis across species: the cases of *Drosophila*, Zebrafish and the mouse. Exp. Cell. Res. 317, 2740-2747.

Friedman, A.A., Tucker, G., Singh, R., Yan, D., Vinayagam, A., Hu, Y., Binari, R., Hong, P., Sun, X., Porto, M., et al., 2011. Proteomic and functional genomic landscape of receptor tyrosine kinase and ras to extracellular signal-regulated kinase signaling. Sci. Signal. 4, rs10.

Friedman, R., 2011. Genomic organization of the glutathione S-transferase family in insects. Mol. Phylogenet. Evol. 61, 924-932.

Garvey, C., Malcolm, C., 2000. *Anopheles stephensi* Dox-A2 shares common ancestry with genes from distant groups of eukaryotes encoding a 26S proteasome subunit and is in a conserved gene cluster. J. Mol. Evol. 50, 497-509.

Ge, X., Kitten, T., Chen, Z., Lee, S.P., Munro, C.L., Xu, P., 2008. Identification of *Streptococcus sanguinis* genes required for biofilm formation and examination of their role in endocarditis virulence. Infect. Immun. 76, 2551-2559.

Gerardo, N.M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., de Vos, M., Duncan, E.J., Evans, J.D., Gabaldon, T., Ghanim, M., et al., 2010. Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. Genome Biol. 11, R21.

Gesellchen, V., Kuttenkeuler, D., Steckel, M., Boutros, M., 2005. An RNA interference

screen identifies Inhibitor of Apoptosis Protein 2 as a regulator of innate immune signalling, in *Drosophila*. EMBO Rep. 6: 979.

Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N., et al., 2011. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 7, e1001393.

Gil, R., Silva, F.J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Hölldobler, B., 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. USA 100, 9388-9393.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al., 2003. A protein interaction map of *Drosophila melanogaster*. Science 302, 1727-1736.

Gorman, M.J., An, C., Kanost, M.R., 2007. Characterization of tyrosine hydroxylase from Manduca sexta. Insect Biochem. Mol. Biol. 37, 1327-1337.

Gottar, M., Gobert, V., Matskevich, A.A., Reichhart, J.-M., Wang, C., Butt, T.M., Belvin, M., Hoffmann, J.A., Ferrandon, D., 2006. Dual detection of fungal infections in *Drosophila* via recognition of glucans and sensing of virulence factors. Cell 127, 1425-1437.

Guntermann, S., Foley, E., 2011. The protein Dredd is an essential component of the c-Jun N-terminal kinase pathway in the *Drosophila* immune response. J. Biol. Chem. 286, 30284-30294.

Gupta, S.K., Bencurova, E., Srivastava, M., Pahlavan, P., Balkenhol, J., Dandekar, T., 2016. Improving Re-annotation of Annotated Eukaryotic Genomes. Big Data Analytics in Genomics., in: Wong, K.C. (Ed.), Big Data Analytics in Genomics. Springer, New York.

Gupta, S.K., Kupper, M., Ratzka, C., Feldhaar, H., Vilcinskas, A., Gross, R., Dandekar, T., Forster, F., 2015. Scrutinizing the immune defence inventory of *Camponotus floridanus* applying total transcriptome sequencing. BMC Genomics 16, 540.

Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al., 2011. A protein complex network of *Drosophila melanogaster*. Cell 147, 690-703.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720-1730.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al., 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654-5666.

Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., Salzberg, S.L., 2002. Full-length messenger RNA sequences greatly improve genome. Genome Biol. 3, 1-0029.0012.

Haglund, C.M., Welch, M.D., 2011. Pathogens and polymers: microbe–host interactions illuminate the cytoskeleton. J. Cell Biol. 195, 7-17.

Haine, E.R., Moret, Y., Siva-Jothy, M.T., Rolff, J., 2008. Antimicrobial Defense and Persistent Infection in Insects. Science 322, 1257-1259.

Hajishengallis, G., Genco, R.J., 2004. Downregulation of the DNA-binding activity of nuclear factor-κB p65 subunit in *Porphyromonas gingivalis* fimbria-induced tolerance. Infect. Immun. 72, 1188-1191.

Hamilton, C., Lejeune, B.T., Rosengaus, R.B., 2010. Trophallaxis and prophylaxis: social immunity in the carpenter ant *Camponotus pennsylvanicus*. Biol. Lett., rsbl20100466.

Hasdemir, D.E., Broemer, M., Leulier, F., Lane, W.S., Paquette, N.P., Hwang, D., Kim, C.-H., Stoven, S., Meier, P., Silverman, N.S., 2009. Two roles for the *Drosophila* IKK complex in the activation of Relish and the induction of antimicrobial

peptide genes. Proc. Natl. Acad. Sci. USA 106, 9779-9784.

Hejazi, A., Falkiner, F., 1997. *Serratia marcescens*. J. Med. Microbiol. 46, 903-912.

Hertle, R., Hilger, M., Weingardt-Kocher, S., Walev, I., 1999. Cytotoxic action of *Serratia marcescens* hemolysin on human epithelial cells. Infect. Immun. 67, 817-825.

Herz, J., Strickland, D.K., 2001. LRP: a multifunctional scavenger and signaling receptor. J. Clin. Invest. 108, 779.

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

Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 1.

Hsu, T., Hutto, D.L., Minion, F.C., Zuerner, R.L., Wannemuehler, M.J., 2001. Cloning of a Beta-hemolysin gene of Brachyspira (Serpulina) hyodysenteriae and its expression in *Escherichia coli*. Infect. Immun. 69, 706-711.

Hughes, A.L., 1999. Evolution of the arthropod prophenoloxidase/hexamerin protein family. Immunogenetics 49, 106-114.

Janeway, C.A., Medzhitov, R., 2002. Innate immune recognition. Annu Rev Immunol 20, 197-216.

Janssen, B.J., Huizinga, E.G., Raaijmakers, H.C., Roos, A., Daha, M.R., Nilsson-Ekdahl, K., Nilsson, B., Gros, P., 2005. Structures of complement component C3 provide insights into the function and evolution of immunity. Nature 437, 505-511.

Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41-42.

Jiang, H., Patel, P.H., Kohlmaier, A., Grenley, M.O., McEwen, D.G., Edgar, B.A., 2009. Cytokine/Jak/Stat signaling mediates regeneration and homeostasis in the *Drosophila* midgut. Cell 137, 1343-1355.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Res. 110, 462-467.

Kajla, M.K., Andreeva, O., Gilbreath, T.M., Paskewitz, S.M., 2010. Characterization of expression, activity and role in antibacterial immunity of *Anopheles gambiae* lysozyme c-1. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 155, 201-209.

Kajla, M.K., Shi, L., Li, B., Luckhart, S., Li, J., Paskewitz, S.M., 2011. A new role for an old antimicrobial: lysozyme c-1 can function to protect malaria parasites in Anopheles mosquitoes. PloS one 6, e19649.

Kallio, J., Leinonen, A., Ulvila, J., Valanne, S., Ezekowitz, R.A., Rämet, M., 2005. Functional analysis of immune response genes in *Drosophila* identifies JNK pathway as a regulator of antimicrobial peptide gene expression in S2 cells. Microbes Infect. 7, 811-819.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27-30.

Keller, O., Odronitz, F., Stanke, M., Kollmar, M., Waack, S., 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9, 278.

Kikutani, H., Kumanogoh, A., 2003. Semaphorins in interactions between T cells and antigen-presenting cells. Nat. Rev. Immunol. 3, 159-167.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.

Kim, S.-H., Lee, W.-J., 2014. Role of DUOX in gut inflammation: lessons from *Drosophila* model of gut-microbiota interactions. Front. Cell. Infect. Microbiol. 3, 116.

Kim, Y., Min, B., Yi, G.-S., 2012. IDDI: integrated domain-domain interaction and

protein interaction analysis system. Proteome Sci. 10, 1.

Kleino, A., Valanne, S., Ulvila, J., Kallio, J., Myllymäki, H., Enwald, H., Stöven, S., Poidevin, M., Ueda, R., Hultmark, D., 2005. Inhibitor of apoptosis 2 and TAK1-binding protein are components of the *Drosophila* Imd pathway. EMBO J. 24, 3423-3434.

Korcsmáros, T., Szalay, M.S., Rovó, P., Palotai, R., Fazekas, D., Lenti, K., Farkas, I.J., Csermely, P., Vellai, T., 2011. Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. PloS one 6, e19240.

Koski, L.B., Gray, M.W., Lang, B.F., Burger, G., 2005. AutoFACT: An Automatic Functional Annotation and Classification Tool. BMC Bioinformatics 6, 1.

Kounatidis, I., Ligoxygakis, P., 2012. *Drosophila* as a model system to unravel the layers of innate immunity to infection. Open Biol. 2, 120075.

Kramer, K.J., Muthukrishnan, S., 1997. Insect chitinases: molecular biology and potential use as biopesticides. Insect Biochem. Mol. Biol. 27, 887-900.

Kremer, N., Charif, D., Henri, H., Gavory, F., Wincker, P., Mavingui, P., Vavre, F., 2012. Influence of *Wolbachia* on host gene expression in an obligatory symbiosis. BMC Microbiol. 12 Suppl 1, S7.

Kuehn, M.J., Kesty, N.C., 2005. Bacterial outer membrane vesicles and the host–pathogen interaction. Genes Dev. 19, 2645-2655.

Kumanogoh, A., Kikutani, H., 2003. Immune semaphorins: a new area of semaphorin research. J. Cell. Sci. 116, 3463-3470.

Kumar, C., Mann, M., 2009. Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett. 583, 1703-1712.

Kumar, R., Nanduri, B., 2010. HPIDB-a unified resource for host-pathogen interactions. BMC Bioinformatics 11, 1.

Labaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S., Kreil, D.P., 2011. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics 27, i383-391.

Lagueux, M., Perrodou, E., Levashina, E.A., Capovilla, M., Hoffmann, J.A., 2000. Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of *Drosophila*. Proc. Natl. Acad. Sci. USA 97, 11427-11432.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al., International Human Genome Sequencing, C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359.

Law, M., Childs, K.L., Campbell, M.S., Stein, J.C., Olson, A.J., Holt, C., Panchy, N., Lei, J., Jiao, D., Andorf, C.M., 2015. Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant Physiol. 167, 25-39.

Lebestky, T., Jung, S.-H., Banerjee, U., 2003. A Serrate-expressing signaling center controls Drosophila hematopoiesis. Genes Dev. 17, 348-353.

Lee, W.-J., 2009. Bacterial-modulated host immunity and stem cell activation for gut homeostasis. Genes Dev. 23, 2260-2265.

Lemaitre, B., Hoffmann, J., 2007. The host defense of *Drosophila melanogaster*. Annu. Rev. Entomol. 25, 697-743.

Leppä, S., Bohmann, D., 1999. Diverse functions of JNK signaling and c-Jun in stress response and apoptosis. Oncogene 18.

Leulier, F., Lemaitre, B., 2008. Toll-like receptors--taking an evolutionary approach. Nat. Rev. Genet. 9, 165-178.

Levashina, E.A., Moita, L.F., Blandin, S., Vriend, G., Lagueux, M., 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.

Lewis, S., Ashburner, M., Reese, M.G., 2000. Annotating eukaryote genomes. Curr. Opin. Struct. Biol. 10, 349-354.

Li, L., Chen, E., Yang, C., Zhu, J., Jayaraman, P., De Pons, J., Kaczorowski, C.C., Jacob, H.J., Greene, A.S., Hodges, M.R., et al., 2015. Improved rat genome gene prediction by integration of ESTs with RNA-Seq information. Bioinformatics 31, 25-32.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178-2189.

Li, S., Jin, X.-K., Guo, X.-N., Yu, A.-Q., Wu, M.-H., Tan, S.-J., Zhu, Y.-T., Li, W.-W., Wang, Q., 2013. A double WAP domain-containing protein Es-DWD1 from *Eriocheir sinensis* exhibits antimicrobial and proteinase inhibitory activities. PloS one 8, e73563.

Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A., Wright, C., Schweitzer, P.A., Gao, Y., et al., 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat. Biotechnol. 32, 915-925.

Li, X., Schuler, M.A., Berenbaum, M.R., 2007. Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu. Rev. Entomol. 52, 231-253.

Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., Lin, K., 2011. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genomics 12, 1.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al., 2012. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40, D857-861.

Liliental, J., Chang, D.D., 1998. Rack1, a receptor for activated protein kinase C, interacts with integrin β subunit. J. Biol. Chem. 273, 2379-2383.

Lin, H.-N., Chen, C.-T., Sung, T.-Y., Ho, S.-Y., Hsu, W.-L., 2009. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. BMC Bioinformatics 10, 1.

Lindmark, H., Johansson, K.C., Stöven, S., Hultmark, D., Engström, Y., Söderhäll, K., 2001. Enteric bacteria counteract lipopolysaccharide induction of antimicrobial peptide genes. J. Immunol. 167, 6920-6923.

Liu, D., Wang, Y., Wei, L., Ye, H., Liu, H., Wang, L., Liu, R., Li, D., Lai, R., 2013. Snake venom-like waprin from the frog of *Ceratophrys calcarata* contains antimicrobial function. Gene 514, 99-104.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012, 251364.

Lourenço, A.P., Martins, J.R., Bitondi, M.M., Simoes, Z.L., 2009. Trade-off between immune stimulation and expression of storage protein genes. Arch. Insect Biochem. Physiol. 71, 70-87.

Lu, A., Zhang, Q., Zhang, J., Yang, B., Wu, K., Xie, W., Luan, Y.-X., Ling, E., 2014. Insect prophenoloxidase: the view beyond immunity. Front. Physiol. 5, 252.

Lukashev, M.E., Werb, Z., 1998. ECM signalling: orchestrating cell behaviour and misbehaviour. Trends Cell. Biol. 8, 437-441.

Luo, Q., Pagel, P., Vilne, B., Frishman, D., 2011. DIMA 3.0: domain interaction map. Nucleic Acids Res. 39, D724-D729.

Marmaras, V.J., Lampropoulou, M., 2009. Regulators and signalling in insect haemocyte immunity. Cell. Signal. 21, 186-195.

Martins, L.M., Iaccarino, I., Tenev, T., Gschmeissner, S., Totty, N.F., Lemoine, N.R., Savopoulos, J., Gray, C.W., Creasy, C.L., Dingwall, C., 2002. The serine protease Omi/HtrA2 regulates apoptosis by binding XIAP through a reaper-like motif. J. Biol. Chem. 277, 439-444.

Maslov, S., Sneppen, K., 2002. Specificity and stability in topology of protein networks. Science 296, 910-913.

Matsumoto, K., Maeda, H., Takata, K., Kamata, R., Okamura, R., 1984. Purification and characterization of four proteases from a clinical isolate of Serratia marcescens kums 3958. J. Bacteriol. 157, 225-232.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., 2009. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 37, D619-D622.

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74, 560-564.

Meinander, A., Runchel, C., Tenev, T., Chen, L., Kim, C.H., Ribeiro, P.S., Broemer, M., Leulier, F., Zvelebil, M., Silverman, N., 2012. Ubiquitylation of the initiator caspase DREDD is required for innate immune signalling. EMBO J. 31, 2770-2783.

Meng, Y., Omuro, N., Funaguma, S., Daimon, T., Kawaoka, S., Katsuma, S., Shimada, T., 2008. Prominent down-regulation of storage protein genes after bacterial challenge in eri-silkworm, *Samia cynthia ricini*. Arch. Insect Biochem. Physiol. 67, 9-19.

Messens, J., Rouhier, N., Collet, J.-F., 2013. Redox homeostasis, Oxidative Stress and Redox Regulation. Springer, pp. 59-84.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al., 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. 3, research0083, 1-22.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res., gkt263.

Mizui, M., Kumanogoh, A., Kikutani, H., 2009. Immune semaphorins: novel features of neural guidance molecules. J. Clin. Immunol. 29, 1-11.

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

Moran, N.A., McCutcheon, J.P., Nakabachi, A., 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42, 165-190.

Moreth, K., Iozzo, R.V., Schaefer, L., 2012. Small leucine-rich proteoglycans orchestrate receptor crosstalk during inflammation. Cell Cycle 11, 2084-2091.

Mosca, T.J., 2015. On the Teneurin track: a new synaptic organization molecule emerges. Front. Cell. Neurosci. 9, 204.

Mostowy, S., Shenoy, A.R., 2015. The cytoskeleton in cell-autonomous immunity: structural determinants of host defence. Nat. Rev. Immunol. 15, 559-573.

Movahedzadeh, F., Williams, A., Clark, S., Hatch, G., Smith, D., ten Bokum, A., Parish, T., Bacon, J., Stoker, N., 2008. Construction of a severely attenuated mutant of *Mycobacterium tuberculosis* for reducing risk to laboratory workers. Tuberculosis 88, 375-381.

Moya, A., Pereto, J., Gil, R., Latorre, A., 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat. Rev. Genet. 9, 218-229.

Mueller, G.A., Edwards, L.L., Aloor, J.J., Fessler, M.B., Glesner, J., Pomés, A., Chapman, M.D., London, R.E., Pedersen, L.C., 2010. The structure of the dust mite allergen Der p 7 reveals similarities to innate immune proteins. J. Allergy Clin. Immunol. 125, 909-917. e904.

Munch, K., Krogh, A., 2006. Automatic generation of gene finders for eukaryotic species. BMC Bioinformatics 7, 1.

Myers, E.W., 1995. Toward simplifying and accurately formulating fragment assembly. J. Comput. Biol. 2, 275-290.

Nagasaka, K., Nakagawa, S., Yano, T., Takizawa, S., Matsumoto, Y., Tsuruga, T., Nakagawa, K., Minaguchi, T., Oda, K.,

Hiraike-Wada, O., 2006. Human homolog of *Drosophila* tumor suppressor Scribble negatively regulates cell-cycle progression from G1 to S phase by localizing at the basolateral membrane in epithelial cells. Cancer Sci. 97, 1217-1225.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314, 267-267.

Neisch, A.L., Speck, O., Stronach, B., Fehon, R.G., 2010. Rho1 regulates apoptosis via activation of the JNK signaling pathway at the plasma membrane. J. Cell Biol. 189, 311-323.

Okazaki, S., Hattori, Y., Saeki, K., 2007. The *Mesorhizobium loti* purB gene is involved in infection thread formation and nodule development in *Lotus japonicus*. J. Bacteriol. 189, 8347-8352.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al., 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42, D358-363.

Osman, D., Buchon, N., Chakrabarti, S., Huang, Y.-T., Su, W.-C., Poidevin, M., Tsai, Y.-C., Lemaitre, B., 2012. Autocrine and paracrine unpaired signaling regulate intestinal stem cell maintenance and division. J. Cell. Sci. 125, 5944-5949.

Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L., 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38, D196-D203.

Otti, O., Tragust, S., Feldhaar, H., 2014. Unifying external and internal immune defences. Trends Ecol. Evol. 29, 625-634.

Papin, J.A., Hunter, T., Palsson, B.O., Subramaniam, S., 2005. Reconstruction of cellular signalling networks and analysis of their properties. Nat. Rev. Mol. Cell Biol. 6, 99-111.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061-1067.

Pereira, C., Denise, A., Lespinet, O., 2014. A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15 Suppl 6, S16.

Perrin, J., Mortier, M., Jacomin, A.-C., Viargues, P., Thevenon, D., Fauvarque, M.-O., 2014. The nonaspanins TM9SF2 and TM9SF4 regulate the plasma membrane localization and signalling activity of the peptidoglycan recognition protein PGRP-LC in *Drosophila*. J. Innate Immun. 7, 37-46.

Phillippy, A.M., Schatz, M.C., Pop, M., 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 9, R55.

Pires-daSilva, A., Sommer, R.J., 2003. The evolution of signalling pathways in animal development. Nat. Rev. Genet. 4, 39-49.

Pompella, A., Visvikis, A., Paolicchi, A., De Tata, V., Casini, A.F., 2003. The changing faces of glutathione, a cellular protagonist. Biochem. Pharmacol. 66, 1499-1503.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10, 354-366.

Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. 24, 142-149.

Purcell, J., Chapuisat, M., 2014. Foster carers influence brood pathogen resistance in ants. Proc. Natl. Acad. Sci. USA 281, 20141338.

Qiu, P., Pan, P.C., Govind, S., 1998. A role for the *Drosophila* Toll/Cactus pathway in larval hematopoiesis. Development 125, 1909-1920.

Raha, S., Robinson, B.H., 2000. Mitochondria, oxygen free radicals, disease and ageing. Trends Biochem. Sci. 25, 502-508.

Randolt, K., Gimple, O., Geissendorfer, J., Reinders, J., Prusko, C., Mueller, M.J., Albert, S., Tautz, J., Beier, H., 2008. Immune-related proteins induced in the hemolymph after aseptic and septic injury

differ in honey bee worker larvae and adults. Arch. Insect Biochem. Physiol. 69, 155-167.

Ranson, H., Rossiter, L., Ortelli, F., Jensen, B., Xuelan, W., Collins, F.H., Hemingway, J., 2001. Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector *Anopheles gambiae*. Biochem. J. 359, 295-304.

Ratzka, C., Forster, F., Liang, C., Kupper, M., Dandekar, T., Feldhaar, H., Gross, R., 2012. Molecular characterization of antimicrobial peptide genes of the carpenter ant *Camponotus floridanus*. PloS one 7, e43036.

Ratzka, C., Gross, R., Feldhaar, H., 2013. Gene expression analysis of the endosymbiont-bearing midgut tissue during ontogeny of the carpenter ant *Camponotus floridanus*. J. Insect Physiol. 59, 611-623.

Ratzka, C., Liang, C.G., Dandekar, T., Gross, R., Feldhaar, H., 2011. Immune response of the ant *Camponotus floridanus* against pathogens and its obligate mutualistic endosymbiont. Insect Biochem. Mol. Biol. 41, 529-536.

Renoz, F., Noël, C., Errachid, A., Foray, V., Hance, T., 2015. Infection dynamic of symbiotic bacteria in the pea aphid *Acyrthosiphon pisum* gut and host immune response at the early steps in the infection process. PloS one 10, e0122099.

Richard, F.-J., Holt, H.L., Grozinger, C.M., 2012. Effects of immunostimulation on social behavior, chemical communication and genome-wide gene expression in honey bee workers (*Apis mellifera*). BMC Genomics 13, 1.

Richard, J.-F., Petit, L., Gibert, M., Marvaud, J.C., Bouchaud, C., Popoff, M.R., 2010. Bacterial toxins modifying the actin cytoskeleton. Int. Microbiol. 2, 185-194.

Rosengaus, R.B., Malak, T., MacKintosh, C., 2013. Immune-priming in ant larvae: social immunity does not undermine individual immunity. Biol. Lett. 9, 20130563.

Roslan, R., Othman, R.M., Shah, Z.A., Kasim, S., Asmuni, H., Taliba, J., Hassan, R., Zakaria, Z., 2010. Utilizing shared interacting domain patterns and Gene Ontology information to improve protein-protein interaction prediction. Comput. Biol. Med. 40, 555-564.

Royet, J., Dziarski, R., 2007. Peptidoglycan recognition proteins: pleiotropic sensors and effectors of antimicrobial defences. Nat. Rev. Microbiol 5, 264-277.

Rutschmann, S., Jung, A.C., Zhou, R., Silverman, N., Hoffmann, J.A., Ferrandon, D., 2000. Role of Drosophila IKKγ in a Toll-independent antibacterial immune response. Nat. Immunol. 1, 342-347.

Sackton, T.B., Werren, J.H., Clark, A.G., 2013. Characterizing the infection-induced transcriptome of *Nasonia vitripennis* reveals a preponderance of taxonomically-restricted immune genes. PloS one 8, e83984.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D., 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, D449-451.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Schaefer, L., Iozzo, R.V., 2008. Biological functions of the small leucine-rich proteoglycans: from genetics to signal transduction. J. Biol. Chem. 283, 21305-21309.

Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20, 1165-1173.

Schlüns, H., Crozier, R., 2007. Relish regulates expression of antimicrobial peptide genes in the honeybee, *Apis mellifera*, shown by RNA interference. Insect Mol. Biol.16, 753-759.

Schmidt, O., Theopold, U., Strand, M., 2001. Innate immunity and its evasion and suppression by hymenopteran endoparasitoids. Bioessays 23, 344-351.

Schwartz, A.S., Yu, J., Gardenour, K.R., Finley, R.L., Jr., Ideker, T., 2009. Cost-effective strategies for completing the interactome. Nat. Methods 6, 55-61.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-2504.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T., 2005. Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. USA 102, 1974-1979.

Shen, Z., 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.

Shi, X.-Z., Zhong, X., Yu, X.-Q., 2012. *Drosophila melanogaster* NPC2 proteins bind bacterial cell wall components and may function in immune signal pathways. Insect Biochem. Mol. Biol. 42, 545-556.

Shin, K., Straight, S., Margolis, B., 2005. PATJ regulates tight junction formation and polarity in mammalian epithelial cells. J. Cell Biol. 168, 705-711.

Shoemaker, B.A., Panchenko, A.R., 2007. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comp. Biol. 3, e43.

Sideri, M., Tsakas, S., Markoutsa, E., Lampropoulou, M., Marmaras, V.J., 2008. Innate immunity in insects: surface-associated dopa decarboxylase-dependent pathways regulate phagocytosis, nodulation and melanization in medfly haemocytes. Immunology 123, 528-537.

Silverman, N., Zhou, R., Erlich, R.L., Hunter, M., Bernstein, E., Schneider, D., Maniatis, T., 2003. Immune activation of NF-κB and JNK requires *Drosophila* TAK1. J. Biol. Chem. 278, 48928-48934.

Skiena, S., 1990. Dijkstra's algorithm. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley, 225-227.

Smit, A., Hubley, R., 2011. RepeatModeler 1.05.repeatmasker.org, http://www.repeatmasker.org/RepeatModeler.html.

Smit, A., Hubley, R., Green, P., 1996-2013. RepeatMasker 3.0 repeatmasker.org, http://www.repeatmasker.org/webrepeatmaskerhelp.html.

Smith, D.A., Parish, T., Stoker, N.G., Bancroft, G.J., 2001. Characterization of auxotrophic mutants of Mycobacterium tuberculosis and their potential as vaccine candidates. Infect. Immun. 69, 1142-1150.

Smith, V.J., 2011. Phylogeny of whey acidic protein (WAP) four-disulfide core proteins and their role in lower vertebrates and invertebrates. Biochem. Soc. Trans. 39, 1403-1408.

Stanyon, C.A., Liu, G., Mangiola, B.A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., Finley, R.L., Jr., 2004. A *Drosophila* protein-interaction map centered on cell-cycle regulators. Genome Biol. 5, R96.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al., 2011. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39, D698-704.

Stoll, S., Feldhaar, H., Fraunholz, M.J., Gross, R., 2010. Bacteriocyte dynamics during development of a holometabolous insect, the carpenter ant *Camponotus floridanus*. BMC Microbiol. 10, 308.

Stoll, S., Feldhaar, H., Gross, R., 2009. Transcriptional profiling of the endosymbiont *Blochmannia floridanus* during different developmental stages of its holometabolous ant host. Environ. Microbiol. 11, 877-888.

Storrs, C.H., Silverstein, S.J., 2007. PATJ, a tight junction-associated PDZ protein, is a novel degradation target of high-risk human papillomavirus E6 and the alternatively spliced isoform 18 E6. J. Virol. 81, 4080-4090.

Strand, M.R., Pech, L.L., 1995. Immunological Basis for Compatibility in

Parasitoid Host Relationships. Annu. Rev. Entomol. 40, 31-56.

Stuart, L.M., Deng, J., Silver, J.M., Takahashi, K., Tseng, A.A., Hennessy, E.J., Ezekowitz, R.A.B., Moore, K.J., 2005. Response to Staphylococcus aureus requires CD36-mediated phagocytosis triggered by the COOH-terminal cytoplasmic domain. J. Cell Biol. 170, 477-485.

Sugiura, T., Sakurai, K., Nagano, Y., 2007. Intracellular characterization of DDX39, a novel growth-associated RNA helicase. Exp. Cell. Res. 313, 782-790.

Tang, H., 2009. Regulation and function of the melanization reaction in Drosophila. Fly 3, 105-111.

Tekir, S.D., Çakır, T., Ardıç, E., Sayılırbaş, A.S., Konuk, G., Konuk, M., Sarıyer, H., Uğurlu, A., Karadeniz, İ., Özgür, A., 2013. PHISTO: pathogen–host interaction search tool. Bioinformatics 29, 1357-1358.

Thoetkiattikul, H., Beck, M.H., Strand, M.R., 2005. Inhibitor κB-like proteins from a polydnavirus inhibit NF-κB activation and suppress the insect immune response. Proc. Natl. Acad. Sci. USA 102, 11426-11431.

Thurston, T.L., Wandel, M.P., von Muhlinen, N., Foeglein, Á., Randow, F., 2012. Galectin 8 targets damaged vesicles for autophagy to defend cells against bacterial invasion. Nature 482, 414-418.

Tian, C., Gao, B., Fang, Q., Ye, G., Zhu, S., 2010. Antimicrobial peptide-like genes in *Nasonia vitripennis*: a genomic perspective. BMC Genomics 11, 1.

Torres, A.M., Wong, H.Y., Desai, M., Moochhala, S., Kuchel, P.W., Kini, R.M., 2003. Identification of a novel family of proteins in snake venoms purification and structural characterization of nawaprin from naja nigricollis snake venom. J. Biol. Chem. 278, 40097-40104.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562-578.

Treangen, T.J., Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36-46.

Tucker, R., Chiquet-Ehrismann, R., 2006. Teneurins: a conserved family of transmembrane proteins involved in intercellular signaling during development. Dev. Biol. 290, 237-245.

Uvell, H., Engstrom, Y., 2007. A multilayered defense against infection: combinatorial control of insect immune genes. Trends Genet. 23, 342-349.

Valanne, S., 2014. Functional genomic analysis of the Drosophila immune response. Dev. Comp. Immunol. 42, 93-101.

Valanne, S., Wang, J.-H., Rämet, M., 2011. The *Drosophila* toll signaling pathway. J. Immunol. 186, 649-656.

Vallet-Gely, I., Lemaitre, B., Boccard, F., 2008. Bacterial strategies to overcome insect defences. Nat. Rev. Microbiol. 6, 302-313.

Van den Bossche, J., Malissen, B., Mantovani, A., De Baetselier, P., Van Ginderachter, J.A., 2012. Regulation and function of the E-cadherin/catenin complex in cells of the monocyte-macrophage lineage and DCs. Blood 119, 1623-1633.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al., 2001. The sequence of the human genome. Science 291, 1304-1351.

Vilcinskas, A., 2013. Evolutionary plasticity of insect immunity. J. Insect Physiol. 59, 123-129.

Vilmos, P., Kurucz, E., 1998. Insect immunity: evolutionary roots of the mammalian innate immune system. Immunol. Lett. 62, 59-66.

Vinayagam, A., Zirin, J., Roesel, C., Hu, Y., Yilmazel, B., Samsonova, A.A., Neumüller, R.A., Mohr, S.E., Perrimon, N., 2014. Integrating protein-protein interaction

networks with phenotypes reveals signs of interactions. Nat. Methods 11, 94-99.

Vizan, J.L., Hernandez-Chico, C., del Castillo, I., Moreno, F., 1991. The peptide antibiotic microcin B17 induces double-strand cleavage of DNA mediated by *E. coli* DNA gyrase. EMBO J. 10, 467.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., Vidal, M., 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 287, 116-122.

Wang, C., Guo, X., Xi, R., 2014. EGFR and Notch signaling respectively regulate proliferative activity and multiple cell lineage differentiation of *Drosophila* gastric stem cells. Cell Res. 24, 610-627.

Wasiluk, K.R., Skubitz, K., Gray, B., 1991. Comparison of granule proteins from human polymorphonuclear leukocytes which are bactericidal toward *Pseudomonas aeruginosa*. Infect. Immun. 59, 4193-4200.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., 2013. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res., gkt1099.

Wilson-Rich, N., Dres, S.T., Starks, P.T., 2008. The ontogeny of immunity: development of innate immune strength in the honey bee (*Apis mellifera*). J. Insect Physiol. 54, 1392-1399.

Wojcik, J., Schachter, V., 2001. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics 17 Suppl 1, S296-305.

Yamamoto, S., Sakai, N., Nakamura, H., Fukagawa, H., Fukuda, K., Takagi, T., 2011. INOH: ontology-based highly structured database of signal transduction pathways. Database 2011, bar052.

Yellaboina, S., Tasneem, A., Zaykin, D.V., Raghavachari, B., Jothi, R., 2011. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res. 39, D730-D735.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M., 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comp. Biol. 3, e59.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., Gerstein, M., 2004. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. Genome Res. 14, 1107-1118.

Zeisel, M.B., Koutsoudakis, G., Schnober, E.K., Haberstroh, A., Blum, H.E., Cosset, F.L., Wakita, T., Jaeck, D., Doffoel, M., Royer, C., 2007. Scavenger receptor class B type I is a key host factor for hepatitis C virus infection required for an entry step closely linked to CD81. Hepatology 46, 1722-1731.

Zhang, L., Morrison, A.J., Thibodeau, P.H., 2015. Interdomain Contacts and the Stability of Serralysin Protease from Serratia marcescens. PloS one 10, e0138419.

Zhang, S., Chen, H., Liu, K., Sun, Z., 2009. Inferring protein function by domain context similarities in protein-protein interaction networks. BMC Bioinformatics 10, 395.

Zhang, Z., Zhu, S., 2012. Comparative genomics analysis of five families of antimicrobial peptide-like genes in seven ant species. Dev. Comp. Immunol. 38, 262-274.

Zhong, J., Zhang, H., Stanyon, C.A., Tromp, G., Finley, R.L., Jr., 2003. A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. Genome Res. 13, 2691-2699.

Zhou, H., Rezaei, J., Hugo, W., Gao, S., Jin, J., Fan, M., Yong, C.H., Wozniak, M., Wong, L., 2013. Stringent DDI-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. BMC Syst. Biol. 7 Suppl 6, S6.

Zhu, Q., Deng, Y., Vanka, P., Brown, S.J., Muthukrishnan, S., Kramer, K.J., 2004. Computational identification of novel chitinase-like proteins in the *Drosophila melanogaster* genome. Bioinformatics 20, 161-169.

Previously Published Material

Some parts of this thesis have been already published in BMC Genomics (Gupta et al., 2015).

List of all sections

Section 3.7: Parts of the text are adapted from the section “Antimicrobial peptides of *C. floridanus* and other hymenoptera” in (Gupta et al., 2015).

Section 3.8: Parts of the text are adapted from the section “Prophenoloxidase, serine proteases and serpins” in (Gupta et al., 2015).

Section 3.9: Parts of the text are adapted from the section “Chitinases, glutathione-S-transferases and nitric oxide synthase (NOS)” in (Gupta et al., 2015).

Section 4.2: Parts of the text are adapted from the sections “Toll signalling pathway”, “IMD and JNK signalling pathways” and “Additional receptor proteins and Jak-Stat signalling pathway” in (Gupta et al., 2015).

Section 4.3: Parts of the text are adapted from the section “Identification of genes differentially expressed (DEGs) after immune challenge” in (Gupta et al., 2015).

Figure 6: The figure has been adapted from "Figure 4" in (Gupta et al., 2015).

Figure 11: The figure has been adapted from "Figure 8" in (Gupta et al., 2015).

Figure 12: The figure has been adapted from "Figure 9" in (Gupta et al., 2015).

Copyright

(Gupta et al., 2015)

List of Figures

List of Tables

List of Publications

Research paper

Gupta SK**,** Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F. Scrutinizing the immune defence inventory of *Camponotus floridanus* applying total transcriptome sequencing. *BMC Genomics* 2015 16:540.

Book chapter

Gupta SK, Bencurova E, Srivastava M, Pahlavan P, Balkenhol J, Dandekar T. Improving Re-annotation of Annotated Eukaryotic Genomes. *Big Data Analytics in Genomics*. Springer 2016.

Review article

Kupper M, Gupta SK, Feldhaar H, Gross R. Versatile roles of the chaperonin GroEL in microorganism-insect interactions. *FEMS Microbiol Lett* 2014 353:1-10.

Note: The results of interactome, signalome, host-pathogen interactions, host-symbiont interactions, miRNA identification and haemolymph mass-spectrometry is currently unpublished.

Other publications [2012-2016; during the time of the thesis involved own work]

Gupta SK, Gross R, Dandekar T. An antibiotic target ranking and prioritization pipeline combining sequence, structure and network-based approaches exemplified for *Serratia marcescens*. *Gene* 2016.

Kaltdorf M, Srivastava M, Gupta SK, Liang C, Binder J, Dietl AM, Meir Z, Haas H, Osherov N, Krappmann S, Dandekar T. Systematic identification of anti-fungal drug targets by a metabolic network approach. *Front Mol Biosci* doi: 10.3389/fmolb.2016.00022.

Liang C, Schaack D, Srivastava M, Gupta SK, Sarukhanyan E, Giese A, Pagels M, Romanov N, Pané-Farré J, Fuchs S, Dandekar T. A *Staphylococcus aureus* proteome Overview: Shared and specific proteins and protein complexes from representative strains of all three clades. *Proteomes* 2016, 4:8.

Gupta SK, Srivastava M, Smita S, Gupta T, Gupta SK. Methods to identify evolutionary conserved regulatory elements using molecular phylogenetics in microbes. *Computational Biology and Bioinformatics: Gene Regulation*. CRC Press. 2016.

Smita S, Singh KP, Akhoon BA, Gupta SK, Gupta SK. Bioinformatics tools for interpretation of data used in molecular identification. *Analyzing Microbes: Manual of Molecular Biology Techniques.* Springer Berlin Heidelberg. 2013, 209-243

Scientific report: Training workshop interdisciplinary life sciences. Akudibillah G, Boas SEM, Carreres BM, Dallinga M, van Dijk AJ, Gupta SK et al. *PeerJ PrePrints* 2014 e654v1.

Singh AK, Kingston JJ, Gupta SK, Batra HV. Recombinant Bivalent Fusion Protein rVE Induces CD4+ and CD8+ T-Cell Mediated Memory Immune Response for Protection Against *Yersinia enterocolitica* Infection. *Front Microbiol* 2015 6:1407.

Singh KP, Verma N, Akhoon BA, Bhatt V, Gupta SK, Gupta SK, Smita S. Sequence-based approach for rapid identification of cross-clade CD8+ T-cell vaccine candidates from all high-risk HPV strains. *3 Biotech* 2016, 6:39.

Gupta SK, Srivastava M, Akhoon BA, Gupta SK, Grabe N. In silico accelerated identification of structurally conserved CD8+ and CD4+ T-cell epitopes in high-risk HPV types. *Infect Genet Evol* 2012, 12:1513-1518.

Akhoon BA, Singh KP, Varshney M, Gupta SK, Shukla Y, Gupta SK. Understanding the mechanism of atovaquone drug resistance in *Plasmodium falciparum* cytochrome b mutation Y268S using computational methods. *PLoS One* 2014 9:e110041

Ranjbar MM, Gupta SK, Ghorban K, Nabian S, Sazmand A, Taheri M, Esfandyari S, Taheri M. Designing and modeling of complex DNA vaccine based on tropomyosin protein of *Boophilus* genus tick. *Appl Biochem Biotechnol* 2015 175:323-39.

Shishir K Gupta

Curriculum Vitae

Personal information

Name	Shishir Kumar Gupta
Date of birth	March 4, 1987
Place of birth	Ballia, India
Nationality	Indian

Education

07/2012-07/2016 **Doctoral study (Bioinformatics)**
Department of Bioinformatics; Department of Microbiology,
Julius-Maximilians-Univerrsity, Würzburg, Germany
Intended degree: PhD (Dr. rer. nat).

2006-2008 **Studies of Bioinformatics**
University Institute of Engineering and Technology,
Chhatrapati Shahu Ji Maharaj University, Kanpur, India
Master of Science; Grade: A; (First Class; 71.3%)

2006-2008 **Studies of Life Science**
Veer Bahadur Singh Purvanchal University, Jaunpur, India
Bachelor of Science; Grade: A; (First Class; 62.3%)

2007 **Dimploma in Programming Techniques and RDBMS**
Software Technology Group (STG) International Limited, Jaunpur, India
Grade: A

Research experience

07/2012-07/2016 **Graduate Research Projects and Doctoral Thesis**
Department of Bioinformatics; Department of Microbiology,
Julius-Maximilians-Univerrsity, Würzburg, Germany

07/2011-07/2012 **GerontoSys**
National Center for Tumor Diseases,
Hamamatsu Tissue Imaging and Analysis (TIGA) Center,
University Hospital Heidelberg, Heidelberg

11/2010-06/2011 **Networking Project (NWP-34), 'Validation of identified screening models for development of new alternative models for evaluation of new drug entities'**
Nanomaterial Toxicology Group, Indian Institute of Toxicology Research, CSIR,
Lucknow, India

11/2008-11/2010 **Ministry of Forest and Environment Project, 'Establishment of ENVIS Center on Plant & Pollution'**
Eco-Auditing Laboratory, National Botanical Research Institute, CSIR,
Lucknow, India

03/2009-02/2011 **Immunoinformatics and Drug designing**
Society for Biological Research & Rural Development, Lucknow, India

Publications

Book chapters

Gupta SK, Bencurova E, Srivastava M, Pahlavan P, Balkenhol J, Dandekar T. Improving Re-annotation of Annotated Eukaryotic Genomes. Big Data Analytics in Genomics. Springer 2016.

Gupta SK, Srivastava M, Smita S, Gupta T, Gupta SK. Methods to identify evolutionary conserved regulatory elements using molecular phylogenetics in microbes. Computational Biology and Bioinformatics: Gene Regulation. CRC Press. 2016.

Smita S, Singh KP, Akhoon BA, **Gupta SK**, Gupta SK. Bioinformatics tools for interpretation of data used in molecular identification. Analyzing Microbes: Manual of Molecular Biology Techniques. Springer Berlin Heidelberg. 2013, 209-243.

Research papers

Gupta SK Gross R, Dandekar T. An antibiotic target ranking and prioritization pipeline combining sequence, structure and network-based approaches exemplified for *Serratia marcescens*. *Gene* 2016.

Kaltdorf M, Srivastava M, **Gupta SK**, Liang C, Binder J, Dietl AM, Meir Z, Haas H, Osherov N, Krappmann S, Dandekar T, 2016. Systematic identification of anti-fungal drug targets by a metabolic network approach. *Front Mol Biosci* 3016, 3, p.22.

Singh KP, Verma N, Akhoon BA, Bhatt V, **Gupta SK**, Gupta SK, Smita S. Sequence-based approach for rapid identification of cross-clade CD8+ T-cell vaccine candidates from all high-risk HPV strains. *3 Biotech* 2016, 6:39.

Liang C, Schaack D, Srivastava M, **Gupta SK**, Sarukhanyan E, Giese A, Pagels M, Romanov N, Pané-Farré J, Fuchs S, Dandekar T. A *Staphylococcus aureus* proteome Overview: Shared and specific proteins and protein complexes from representative strains of all three clades. *Proteomes* 2016, 4:8.

Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F. Scrutinizing the immune defence inventory of *Camponotus floridanus* applying total transcriptome sequencing. *BMC Genomics* 2015 16:540.

Kupper M, **Gupta SK**, Feldhaar H, Gross R. Versatile roles of the chaperonin GroEL in microorganism-insect interactions. *FEMS Microbiol Lett* 2014 353:1-10.

Singh AK, Kingston JJ, **Gupta SK**, Batra HV. Recombinant Bivalent Fusion Protein rVE Induces CD4+ and CD8+ T-Cell Mediated Memory Immune Response for Protection Against *Yersinia enterocolitica* Infection. *Front Microbiol* 2015 6:1407.

Ranjbar MM, **Gupta SK**, Ghorban K, Nabian S, Sazmand A, Taheri M, Esfandyari S, Taheri M. Designing and modeling of complex DNA vaccine based on tropomyosin protein of *Boophilus* genus tick. *Appl Biochem Biotechnol* 2015 175:323-39.

BA Akhoon, KP Singh, M Varshney, **SK Gupta**, Y Shukla, SK Gupta. Understanding the mechanism of atovaquone drug resistance in *Plasmodium falciparum* cytochrome b mutation Y268S using computational methods. *PLoS One* 2014 9:e110041.

Gupta SK, Srivastava M, Akhoon BA, Gupta SK, Grabe N. In silico accelerated identification of structurally conserved CD8+ and CD4+ T-cell epitopes in high-risk HPV types. *Infect Genet Evol* 2012, 12:1513-1518.

Gupta SK, **Gupta SK**, Smita S, Srivastava M, Lai X, Schmitz U, Rahman Q, Wolkenhauer O, Vera J. Computational analysis and modeling the effectiveness of 'Zanamivir' targeting neuraminidase protein in pandemic H1N1 strains. *Infect Genet Evol* 2011, 11:1072-1082.

Gupta SK, **Gupta SK**, Smita S, Srivastava M, Lai X, Schmitz U, Rahman Q, Wolkenhauer O, Vera J. Computational analysis and modeling the effectiveness of 'Zanamivir' targeting neuraminidase protein in pandemic H1N1 strains. *Infect Genet Evol* 2011, 11:1072-1082.

Gupta SK, **Gupta SK**, Smita S, Srivastava M, Lai X, Schmitz U, Rahman Q, Wolkenhauer O, Vera J. Computational analysis and modeling the effectiveness of 'Zanamivir' targeting neuraminidase protein in pandemic H1N1 strains. *Infect Genet Evol* 2011, 11:1072-1082.

Baloria U, Akhoon BA, **Gupta SK**, Sharma S, Verma V. In silico proteomic characterization of human epidermal growth factor receptor 2 (HER- 2) for the mapping of high affinity antigenic determinants against breast cancer. *Amino Acids* 2012, 42:1349-60.

Srivastava M, **Gupta SK**, Abhilash PC, Singh N. Structure prediction and binding sites analysis of curcin protein of *Jatropha curcas* using computational approaches. *J Mol Model* 2011, 18:2971-9.

Akhoon BA, **Gupta SK**, Dhaliwal G, Srivastava M, Gupta SK. Virtual screening of specific chemical compounds by exploring *E. coli* NAD+-dependent DNA ligase as a target for antibacterial drug discovery. *J Mol Model* 2011, 17:265-273.

Akhoon BA, Slathia PS, Sharma P, **Gupta SK**, Verma V. In silico identification of novel protective VSG antigens expressed by *Trypanosoma brucei* and an effort for designing a highly immunogenic DNA vaccine using IL-12 as adjuvant. *Microb Pathog* 2011, 51:77-87.

Gupta SK, Singh A, Srivastava M, Gupta SK, Akhoon BA. In silico DNA vaccine designing against human papillomavirus (HPV) causing cervical cancer. *Vaccine* 2010, 28:120-131.

Gupta SK, Smita S, Sarangi AN, Srivastava M, Akhoon BA, Rahman Q, Gupta SK. In silico CD4+ T-cell epitope prediction and HLA distribution analysis for the potential proteins of *Neisseria meningitidis* Serogroup B—A clue for vaccine development. *Vaccine* 2010, 28:7092–7097.

Srivastava M, Akhoon BA, **Gupta SK**, Gupta SK. Development of resistance against blackleg disease in *Brassica oleracea var. botrytis* through in silico methods. *Fungal Genet Biol* 2010, 47:800–808.

Akhoon BA, **Gupta SK**, Verma V, Dhaliwal G, Srivastava M, Gupta SK, Ahmad RF. In silico designing and optimization of anti-Breast cancer antibody mimetic oligopeptide targeting HER-2 in Women. *J Mol Graph Model* 2010, 28:664-669.

Gupta SK, Akhoon BA, Srivastava M, Gupta SK. A Novel Algorithm to Design an Efficient siRNA by Combining the Pre Proposed Rules of siRNA Designing. *Journal of Computer Science & Systems Biology* 2010, 3:005-009.

Srivastava M, **Gupta SK**, Saxena AP, Shittu LAJ, Gupta SK. A Review of Occurrence of Fungal Pathogens on Significant Brassicaceous Vegetable Crops and their Control Measures. *Asian J Agric Sci*, 2010, 2:70-79.

Gupta SK, Singh A, Srivastava M. Designing of drug for removal of methylation from Tcf21 gene in lung cancer. *Online J Bioinform* 2009, 11:149-155.

Conference/Workshop publications

Scientific report: Training workshop interdisciplinary life sciences. Akudibillah G, Boas SEM, Carreres BM, Dallinga M, van Dijk AJ, **Gupta SK** et al. PeerJ PrePrints 2014 e654v1.

Gupta SK, Baweja L, Gurbani D, Pandey AK, Dhawan A. Interaction of C60 fullerene with the proteins involved in DNA mismatch repair pathway. *J Biomed Nanotech* 2011, 7(1):179-180.

Presentations

Talk

An introduction to Systems Biology, Computational tools in Microbial Research NBAIM, Mau, India, November 19th - 30th 2013.

Poster presentations

Re-annotation of the ant *Camponotus floridanus* genome, comprehensive analysis of its immune transcriptome and general reconstruction of ant interactomes (**Gupta SK**, Gross R, Dandekar T.), ECCB 2016, Hague, The Netherlands, September 3-7, 2016.

Transcriptome based re-annotation and comprehensive analysis of immuno-transcriptome of carpenter ant *Camponotus floridanus* (**Gupta SK**, Gross R, Dandekar T.), Ants 2016: Ants and their biotic environment, Munich, Germany, May 5-7, 2016.

How essential are non-essential genes in insect endosymbiont *Blochmannia floridanus*? (**Gupta SK**, Gross R, Dandekar T.), 3rd Mol Micro Meeting, IMIB Würzburg, Germany, May 7-9, 2014.

A combined 3D tissue and in silico tumor model for signaling and metabolism analysis in drug testing (Wangorsch G, Hoff N, Stratmann AT, Kunz M, **Gupta SK**, Göttlich C, Walles H, Nietzer SL, Dandekar T, Dandekar G), Cancer and Metabolism, Amsterdam, The Netherlands, June 24-25, 2013.

Interaction of C60 fullerene with the proteins involved in DNA mismatch repair pathway (**Gupta SK**, Baweja L, Gurbani D, Pandey AK, Dhawan), International Symposium on the Safe Use of Nanomaterials and Workshop on Nanomaterial Safety: Status, Procedures, Policy and Ethical Concerns, IITR, Lucknow, India, 01-03 February, 2011.

Analysis of metal binding sites of Cation Diffusion Facilitator protein to enhance the phytoremediation efficiency of plants" (Srivastava M, **Gupta SK**, Singh N), 4th International conference of Plants and Environmental Pollution, NBRI, Lucknow, India, 8-11 December, 2010.

In-silico designing and optimization of anti-Breast cancer antibody mimetic oligopeptide targeting HER-2 in Women (Akhoon BA, **Gupta SK**, Verma V, Dhaliwal G, Srivastava M, Gupta SK, Ahmed RF), 1st IFIP International conference on Bioinformatics, SVNIT, Surat, India, March 25-28, 2010.