5-2009

COMPARATIVE GENOMICS AND MOLECULAR EVOLUTION: NEW GENOMIC RESOURCES FOR THE HYMENOPTERA AND EVOLUTIONARY STUDIES ON THE GENES OF THE *Nasonia vitripennis* HOX COMPLEX.

Monica Munoz-torres
Clemson University, monica.cecilia@gmail.com

Recommended Citation

Munoz-torres, Monica, "COMPARATIVE GENOMICS AND MOLECULAR EVOLUTION: NEW GENOMIC RESOURCES FOR THE HYMENOPTERA AND EVOLUTIONARY STUDIES ON THE GENES OF THE *Nasonia vitripennis* HOX COMPLEX." (2009). *All Dissertations*. 342.
https://tigerprints.clemson.edu/all_dissertations/342

COMPARATIVE GENOMICS AND MOLECULAR EVOLUTION: NEW GENOMIC RESOURCES FOR THE HYMENOPTERA AND EVOLUTIONARY STUDIES ON THE GENES OF THE *Nasonia vitripennis* HOX COMPLEX.

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Genetics

by
Mónica Cecilia Muñoz-Torres
May 2009

Accepted by:
Amy Lawton-Rauh, Committee Co-chair
Richard Hilderman, Committee Co-chair
David G Heckel
James C Morris
Matthew W Turnbull

ABSTRACT

Research on insects, the most successful group from all metazoans on earth, has important societal, as well as scientific benefits. Insects occupy a wide range of roles, which have an effect on human life either because the former pose serious threats to public health and commercial crops as well as in some cases represent the only way to propagate food resources. Despite their tremendous importance, insect genomics remained an uneven territory dominated by studies in the Drosophila group and the mosquitoes. This dissertation attempts to: 1) report on advances in the development and characterization of genomic tools for species of the order Hymenoptera in the hopes of helping to close this gap; and 2) to shed light on the organization, origin and evolution of genes of the *Hox* cluster in species of the order Hymenoptera through molecular evolution analyses that were possible thanks to the availability of the aforementioned genomic resources.

DEDICATION

To my father, who lost three of his girls to foreign lands and never got us back. And to my mother, who got her gray hair worrying about my father's shrinking heart.

A papi, quien perdió a tres de sus niñas en tierras lejanas y nunca nos volvió a tener a su lado. Y a mami, quien se ganó todas sus canas preocupándose por el corazón arrugadito de papá.

ACKNOWLEDGMENTS

There are many, many people without whom this accomplishment could have not been possible. My eternal gratitude goes to: God. My parents Hernando and Emperatriz, and my siblings Tulia, Isabel and Victoria, for their everlasting love and support. Maria and Morris Levy and Mike and Kathy Panther, my family and home away from home. Adam Cross, a rock upon which I was able to rebuild myself and remain strong. Fred Mercadante, Fr. Emmanuel Andinam and St. Andrew's Church for feeding my soul. Dee Cohn, Lina Ramirez-Lopez and Mónica Martuscielo-Dakunchack, my sisters. Chris Saski, my brother and auxiliary brain. Meg Staton, my wailing wall, box of tissues, and fountain of daily laughs and endless encouragement. Dr. Amy Lawton-Rauh, for her incredible and unconditional support. For her advice. For stepping in, becoming the real advisor. Dean Bruce Rafert, for putting his confidence in me and for his enormous support. Vice President Gail DiSabatino and Associate Dean Rusty Guill, for listening. Dr. David Heckel, Dr. James Morris, and Dr. Matthew Turnbull, members of my graduate committee. Dr. Jack Werren. One million pages would not be enough to express why I am eternally grateful with Jack. I hope this simple "*Thank you*" does it for now. Dr. Chris Mathews -and girls-, for the music. Paul and Evelyn Abenante, Alex Engau, Carla Mora, Jeremy Thompson, Jorge Mogollón, Sarah DiFurio, Angela Saski, Frank Barretta, Miriam Rosemberg, Raewyn Lowe, Megan Ladd, CJ Geraci, Ian, Stephanie and Euan Stocks, Darío Correa, Sacha Lozano and Pili Chaves, Traci Pawlowski, Beto Lippert, Jim and Elizabeth Woodruff and Jon Luttrell, for their unconditional friendship, love and support.

TABLE OF CONTENTS

Page

Table of Contents (Continued)

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

PREFACE

Arthropods are the most diverse group of organisms on earth. With almost one million described species, insects outnumber all other known animal species. Their morphological complexity and developmental diversity have led to major breakthroughs in our understanding of ontogeny and developmental biology, even for distantly related organisms. The wide variety of extant insect forms and life styles has also made them subjects of phylogenetic, evolutionary and ecological studies. Research on insects has important societal, as well as scientific benefits, as many insects pose serious threats to public health and commercial crops.

Despite their tremendous importance, insect genomics remained an uneven territory dominated by studies in the Drosophila group and the mosquitoes. Other than the well-developed genetic model *Drosophila melanogaster*, genomic resources for the remainder of the insects are surprisingly inadequate. EST libraries and genetic linkage maps have been constructed for several insects, but Bacterial Artificial Chromosome (BAC) libraries are available for only a few species such as the honey bee *Apis mellifera* (Tomkins et al., 2002), the mosquitoes *Aedes aegypti* (Jiménez et al., 2004) and *Anopheles gambiae* (Hong et al., 2003), the red flour beetle *Tribolium castaneum* (Brown et al., 2002) and the silk moth *Bombyx mori* (Mita et al., 2002).

The following pages detail research conducted in an attempt to help closing the gap in insect genomics. With this idea in mind, the goal of this dissertation is twofold. First, is to report on advances in the development and characterization of genomic tools for species of the order Hymenoptera in the form of BAC libraries for two species of parasitic wasps of the genus *Nasonia*, and one species of social Hymenoptera, the bumblebee *Bombus terrestris*. The availability of genomic resources will greatly facilitate comparative genomics and positional cloning projects between species of this order. Second, is to shed light on the organization, origin and evolution of genes of the *Hox* cluster in species of the order Hymenoptera through molecular evolution analyses, demonstrating also the utility of developing genomic resources of this kind.

To highlight the most important points for each objective this dissertation is organized into one chapter of literature review followed by three research chapters, which are divided into two sections. The literature review presents a description of the most recent advances in the field of insect genomics, and discusses critical concepts on the theory and implementation of comparative analyses at the molecular level using genomic information.

Following this review, there are two sections of research chapters. The first section includes chapters two and three, and is focused on the generation of genomic resources for the Hymenoptera research community. Chapter Two describes the construction of a publicly available BAC library for the bumblebee (*Bombus terrestris*) and it was published in *Insectes Sociaux*. [Wilfert, L, M Muñoz Torres, C Reber-Funk, R Schmid-Hempel, J Tomkins, J Gadau and P Schmid-Hempel. 2008. Construction and characterization of a BAC-library for a key pollinator, the bumblebee *Bombus terrestris* L. DOI: 10.1007/s00040-008-1034-1]. Bumblebee, a primitive social hymenopteran, is an ecological and evolutionary model species as well as an important agricultural pollinator

(Goulson, 2003). Recent studies on bumblebees have focused on ecological immunity and host-parasite interactions. Analysis of quantitative trait loci (QTL) yielded useful information to study the maintenance of genetic variation for fitness-relevant traits involved in host-parasite interactions in natural populations (Schmid-Hempel, 2001). However it is necessary to rely on the development of genomic resources to test the hypotheses generated through these studies by isolating the genes responsible for the observed variation in the studied populations regarding host-resistance. The development of large-insert libraries open the possibility to resolve this and similar situations in which traditional genetic studies have been exhausted and the need for a concrete answer is necessary in the absence of a whole genome sequence.

Chapter Three is a contribution to a collection of companion papers to be published along with the first assembly of the genome sequence of the parasitic wasp *Nasonia vitripennis*. It constituted the first genome-wide survey of two species of the genus *Nasonia*, *N. vitripennis* and *N. giraulti*, through the construction, characterization and hybridization of BAC libraries. The completed manuscript will be submitted to *Insect Molecular Biology*. [Munoz-Torres, M, C Saski, B Blackmon, J Romero-Severson, J Tomkins and JH Werren. Development of BAC library resources for parasitic Hymenoptera: (*Nasonia vitripennis* and *Nasonia giraulti.* Pteromalidae). *Insect Molecular Biology*. In preparation]. Laboratory tractability, interesting and diverse biology, large family sizes, haplodiploidy, the ability to inbreed and produce healthy isogenic inbred lines, a wealth of visible and molecular markers, four closely related and interfertile species, the ease of performing complete genome screenings in the search for mutations in the haploid sex and the capacity to produce genetically identical recombinant genotypes in the F3 generation have made the species of the genus *Nasonia* a primary model for parasitoid genetics (e.g. Werren, 2000, Gadau et al., 1999, Darling and Werren, 1990,Whiting, 1967). More is known about the biology of *Nasonia* than any other parasitic hymenopteran, and the species have a very active research community dedicated to study a diverse collection of aspects of its natural history, development, ecology, morphology, physiology and genetics, to name a few. Yet until recently, very little was known about the structure of the genome of *Nasonia*, and genomic resources to answer specific questions were scarce. This chapter describes a gateway into the genome of two species of *Nasonia*.

The second section, chapter four, describes molecular and evolutionary analyses of *Antennapedia* (*Antp*) in the context of patterns observed among eight of the ten orthologs of the insect *Hox* cluster in *Nasonia vitripennis.* The experimental approach and results presented test hypotheses regarding nucleotide substitution patterns expected during alternative evolutionary processes. This chapter focuses on substitution rates to estimate divergence within the *Antennapedia* (*Antp*) gene among three species of Hymenoptera, and investigated the divergence among genes of the *Hox* cluster of genes of *Nasonia vitripennis* as a comparison to explain the observed patterns in the *N. vitripennis Antp*-gene nucleotide sequence. The manuscript for this chapter was prepared for submission to *Insect Molecular Biology*. [Muñoz-Torres, M and A Lawton-Rauh. Selection differential across *Antennapedia* among hymenopteran species suggest homeodomain-specific purifying selection. *Insect Molecular Biology*. In preparation]. Appendix C contains supplementary data in support of this chapter.

Events of gene transfers between organisms of different species provide yet a different perspective in studying the origin and evolution of genes; this equation becomes a bit more complex when the two organisms in question belong to distant branches of the evolutionary tree. Research reporting on widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes showed that some of the inserted bacterial genes are transcribed within eukaryotic cells (Dunning Hotopp et al., 2007); this suggests that these heritable lateral gene transfers may provide a mechanism for acquisition of new genes and functions. Which are these mechanisms? How do they drive gene evolution? How have both eukaryotic machinery and genes (and the genomic vicinity) changed since the bacterial insertion event? How is selection acting on the inserted DNA? Do endosymbiotic bacterial accelerate divergence and speciation of their eukaryotic host? These and many other questions regarding issues such as transmission of the inserted DNA and how endosymbionts, located in the germlines of their eukaryotic hosts, manipulate host cell biology and reproduction, still remain to be answered.

Appendix A contains the published manuscript on evidence obtained for events of lateral gene transfer from *Wolbachia* to 11 different species of Eukaryotes. [Dunning Hotopp, JC, ME Clark, DCSG Oliveira, JM Foster, P Fischer, MC Muñoz Torres, JD Giebel, N Kumar, N Ishmael, S Wang, J Ingram, RV Nene, J Shepard, J Tomkins, S Richards, DJ Spiro, E Ghedin, BE Slatko, H Tettelin, and JH Werren. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science*

317 (5845):1753-1756.]. Appendix B contains data concerning ongoing research to help answer these and other questions related to a lateral gene transfer event reported from endosymbiotic *Wolbachia pipientis* into the genome of the tropical fruit fly *Drosophila ananassae*. The appendix describes results of hybridization experiments with *Wolbachia*-specific genes on a *Drosophila ananassae* (Hawai'i) BAC library. The BAC library was constructed with adult individuals of a non-cured strain, thus BACs may contain endosymbiont DNA or *Wolbachia*-inserted DNA. Using Finger Printed Contig (FPC) analyses and BAC-end sequence data, seven candidate BACs were chosen for full sequencing.

References

Brown, S, JP Fellersb, TD Shippy, EA Richardson, M Maxwell, JJ Stuart, and RE Denella. 2002. Sequence of the *Tribolium castaneum* Homeotic Complex: The Region Corresponding to the Drosophila melanogaster Antennapedia Complex. *Genetics* **160**: 1067-1074.

Darling, DC and JH Werren. 1990. Biosystems of *Nasonia* (Hymenoptera: Pteromalidae): two new species reared from bird's nests in North America. *Ann Ent Soc Am* **83**:352-370.

Dunning Hotopp, JC, ME Clark, DCSG Oliveira, JM Foster, P Fischer, MC Muñoz Torres, JD Giebel, N Kumar, N Ishmael, S Wang, J Ingram, RV Nene, J Shepard, J Tomkins, S Richards, DJ Spiro, E Ghedin, BE Slatko, H Tettelin, and JH Werren. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science* **317**(5845):1753-1756.

Gadau, J, RE Page Jr. and JH Werren. 1999. Mapping of Hybrid Incompatibility in *Nasonia*. *Genetics* **153**:1731-1741.

Goulson D. 2003. Bumblebees - their Behaviour and Ecology. Oxford University Press, New York 235p.

Hong, YS, JR Hogan, X Wang, A Sarkar, C Sim, BJ Loftus, C Ren, ER Huff, JL Carlile, K Black, HB Zhang, MJ Gardner and FH Collins. 2003. Construction of a BAC library and generation of BAC end sequences-tagged connectors for genome sequencing of the African malaria mosquito *Anopheles gambiae*. *Mol Gen Genomics* **268**:720-728.

Jiménez, LV, BK Kang, B de Bruyn, DD Lovin and DW Severson. 2004. Characterization of an *Aedes aegypti* bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence Plasmodium susceptibility. *Insect Mol Biol* **13**(1):37-44.

Mita, K, M Mitsuoki, O Kazuhiro, K Yoshiko, N Junko, MG Suzuki and T Shimada. 2002. Construction of an EST database for Bombyx mori and its applications. *Current Science* **83**(4):426-431.

Schmid-Hempel P. 2001. On the evolutionary ecology of host-parasite interactions: addressing the question with regard to bumblebees and their parasites. *Naturwissenschaften* **88**:147-158.

Tomkins, J, M Luo, G Fang, D Main, J Goicoechea, M Atkins, D Frisch, RE Page, E Guzman-Novoa, Y Yu, GJ Hunt and R Wing. 2002. New genomic resources for the honey bee (*Apis mellifera* L.): development of a deep-coverage BAC library and a preliminary STC database. *Genet Mol Res* **1**:306-316.

Werren, JH. 2000. *Nasonia*: An ideal organism for research and training. Internal document of the Department of Biology. University of Rochester. 17p.

Whiting, AR. 1967. The biology of the parasitic wasp *Mormoniella vitripennis*. [=*Nasonia brevicornis*] (Walker). *Q Rev Biol* **42**:333-406.

CHAPTER ONE

LITERATURE REVIEW ON INSECT COMPARATIVE GENOMICS AND MOLECULAR EVOLUTION

Genomics: definition and use

Dr. Thomas H. Roderick coined the word *genomics* in 1986[1] (Kuska, 1998). At the time Roderick was a geneticist at the Jackson Laboratory in Bar Harbor, Maine and what was initially thought as the name of a journal dedicated to advancements in genome sequencing and mapping, would go on to become a very important part of biology. Today it is widely accepted that the overarching aim of *genomics* is to provide a comprehensive, genome-level understanding of the molecular basis of the structure, functions, and evolution of biological systems using whole-genome sequence information and high-throughput genomic technologies (Zhou et al., 2004).

When whole-genome sequences are not yet available for the organism of interest, it is necessary to resource to tools that will allow researchers to interrogate these genomes. Pursuing a wider understanding of evolution and development, and striving to provide stronger genetic foundations to research on the species of Insecta, genomic resources were developed for a number of insect taxa in recent years using the Bacterial Artificial Chromosome (BAC) system. Libraries were developed for a variety of representative species spanning a large evolutionary distance in the phylogeny of insects, including honey bee *Apis mellifera* (Tomkins et al., 2002), the mosquitoes *Aedes aegypti*

[1] While attending a meeting on "the feasibility of mapping the entire human genome". The word was Roderick's contribution to a discussion on the name of a new genome-oriented scientific journal. It stuck.

(Jiménez et al., 2004) and *Anopheles gambiae* (Hong et al., 2003), the red flour beetle *Tribolium castaneum* (Brown et al., 2002) and the silk moth *Bombyx mori* (Mita et al., 2002). These species shared both the fact that they display attributes that have granted them scientific importance as model organisms for ecological or developmental studies, as well as the scarcity of existing genomic resources at the time. In combination with then extant whole genome sequences of *Drosophila melanogaster*, *Anopheles gambiae* and *Apis mellifera*, such libraries enabled the scientific community to test important hypotheses concerning insect genetics, genomics, development, ecology, systematics, and evolution.

BAC libraries originated as a system to facilitate the construction of large-insert DNA collections of complex genomes with fuller representation and subsequent rapid analysis of complex genomic structure (Shizuya et al., 1992). BACs are not artificial chromosomes per se, but rather are modified bacterial F factors. Although they can carry inserts approaching 500 Kb in length, insert sizes between 80 and 300 kb are more typical (Budiman et al., 2000, Koike et al., 2003, Yu et al., 2000). BACs are relatively free of the chimerism and insert rearrangements that commonly occur in YACs (Woo et al., 1994; Boysen et al., 1997; Cai et al., 1995). BAC clones are relatively easy to manipulate and propagate, thus BAC libraries in which each clone is stored and archived individually (i.e., ordered libraries) have become a central tool in modern genomics research. The library construction process may introduce certain biases that could jeopardize the inclusion of all genomic regions in a single library. There may be overrepresentation and underrepresentation of certain regions of the genome due to the use of a single restriction enzyme per library, partial digestion of genomic DNA, and the unavailability of certain regions of genomes such as centromeres, highly repetitive sequences and telomeres, due to their lack of recognition sites for common restriction enzymes (Mahtani and Willard, 1998; Chew et al., 2002; Yuan et al., 2008). To ensure sufficient depth of genome coverage, attempts are made to redundantly represent the genome by robotically picking recombinant colonies in numbers up to ten times the minimum amount necessary to cover once the size of the genome. Despite these efforts gaps may still exist and alternative tools may be needed to reach the missing regions of the genome. Osegawa and colleagues (2001) and the Lucigen Corporation (2008) developed shearing techniques to partition megabase-sized DNA onto fragments, which can be then cloned onto BAC or fosmid vectors. Mechanical shearing of genomic DNA allows access to all portions of the genome regardless of their availability of restriction sites for the most commonly used cloning enzymes, avoiding bias of aberrant representation of certain regions and increasing genome coverage.

Genomic research is comparative by nature

The discovery of Mendel's laws of heredity developed into a scientific quest to understand the nature and content of genetic information over the last century. The field of biology emerged as a discipline rooted in comparisons. Comparative physiology has assembled a detailed catalogue of the biological similarities and differences between species, revealing insights as to how life has adapted to fill a wide range of environmental niches (Nobrega and Pennacchio, 2003). Comparative studies in anatomy, biochemistry, pharmacology, and immunology and cell biology have provided fundamental paradigms, which have contributed to the growth of each of these disciplines. Genomics is the most recent branch of biology to employ comparison-based strategies (Nobrega and Pennacchio, 2003). Comparative genomics provides a powerful and general approach for indentifying functional elements without previous biological knowledge. It aids in the identification of genes, gene structure, regulatory elements, and evolutionary forces acting on an organism's biological processes surmounting the need for survival. Attaining this information might finally bring us closer to understanding how species have managed to maintaining entire sets of genes which are conserved among many species, as well as developing sets of genes unique to each of them.

Comparative genomics analyses have lead to changes in our understanding of some phylogenetic relationships along the branches of the evolutionary tree. The beetle species richness is now better explained by high survival of lineages and sustained diversification in a variety of niches (Hunt, 2007) and evidence has now surfaced placing the species of the order Hymenoptera at the base of the radiation of Holometabolous insects (Zdobnov and Bork, 2006; Savard et al., 2006). In light of the possibilities offered by the availability of whole-genome sequences of three vector mosquitoes (*Anopheles gambiae, Aedes aegypti* and *Culex pipiens*) comparative genomics also helped researchers to study the mechanisms by which viruses are able to circumvent their host's antiviral interference RNAi, as well as the limitations of this defense (Campbell et al., 2008). Selective pressure exists on both the virus and the vector mosquito to modulate the immune response. The authors reported that in all three species of mosquito anti-viral defense effectors are evolving at a faster rate than those involved in housekeeping functions. Despite this some mosquito species are still effective transmitters of aborviruses regardless of the presence of an anti-viral response, suggesting that both vector and virus are continuously evolving to overcome the challenges posed by the biology of the other. Campbell and colleagues also reported that analysis of genes of the Argonaute protein family (involved in gene silencing pathways through the use of dsRNA) suggested that the small regulatory RNA pathways of *A. aegypti* and *C. pipiens* are evolving faster than those of *A. gambiae* and *D. melanogaster*. Further comparative genomics and functional analyses may be the key to understanding why *A. gambiae* displays a lower competence as a viral vector (it primarily transmits malaria parasites) in what appears to be a more effective antiviral immune response than either of the former species.

Long, continuous segments of DNA and all characterization data available, including their association with chromosomes in our organism of interest, sit at the core of genomic research approaches to understanding evolutionary biology and phylogenetic relationships with respect to other species through comparative analysis. This clear necessity, and the successful stories reported for many animal and plant species, has increased our interest in developing genomic tools for a number of organisms. Currently such tools are being developed primarily in an attempt to understand more about the mechanisms of genome evolution, to better understand our own genome and those of model species involved in research concerning human health and agricultural practices.

Insect genomics

In 2007 twenty-seven insect species had either been recently sequenced or were in the process of being sequenced (Grimmelikhuijzen et al., 2007). Currently, the National Center for Biotechnology Information (NCBI) lists a total of 59 public insect genome-sequencing projects (www.ncbi.nlm.nih.gov/Genomes). There are at least preliminary assembly versions available for 31 species, such as the jewel wasp (*Nasonia vitripennis)* and the mosquito vector of lymphatic filariasis (*Culex quinquefasciatus)*, and in some cases assemblies are available in revised versions (e.g. *Apis mellifera* Assembly v 4.0). To date the genome of the fruit fly *Drosophila melanogaster* is the most thoroughly annotated sequence and contains no gaps. The genomes of 27 additional insect species are still being sequenced and pending for a first assembly version. The list of insect projects currently undergoing covers an evolutionary distance of 310 - 325 millions of years (My) across six orders (Grimaldi and Engel, 2005). Overall, the outcome of these projects will provide genomic information immediately valuable to biochemists, molecular biologists, insect physiologists, and to many other disciplines on the long run. The findings obtained as a result of these genome projects will also very likely uncover a number of surprises such as smaller than expected genome sizes for a number of species (Johnston et al., 2007) or larger than expected numbers of paralogs in collections of insect-specific proteins, which may indicate adaptation to environment changes (Zhang et al., 2007).

Based on their mode of development, insect species may be grouped into non-holometabolous lineages and a monophyletic group that exhibits a holometabolous mode of development as one of their synapomorphies, the Holometabola (Endopterygota) (Gullan and Cranston, 2004). Within non-holometabolous lineages, there are species with a hemimetabolous mode of development. In hemimetabolous insects early developmental stages resemble the adult forms and in holometabolous complete metamorphosis changes the appearance of the immature insect to adopt a fully developed adult form. Intricacies of the discussion defining the difference between hemimetabolous insects and other non-holometabolous developmental schemes go beyond the scope of this review; additionally, all species of non-holometabolous insects included in this review do observe a hemimetabolous mode of development, thus the term hemimetabolous will be used hereafter to refer to all non-holometabolous insect species in this document.

Genome sequencing efforts were initially concentrated on holometabolous insects but in the last few years an increasing number of hemimetabolous insect species have been selected for sequencing. Including species less derived than the predominantly holometabolous ones to the wealth of insect genome projects provides a source of important information for the advancement of research focused on the evolution of insect and arthropod genomes. Their characterization may provide insights into the implications of the enormous amounts of extra genomic DNA on chromosomal structure and organization, and on other basic biological aspects such as mitosis and meiosis.

Public Genome Projects on Species of Holometabolous Insects.

Their considerable impact on human health as vectors of infectious and sometimes fatal diseases, on human agricultural practices as the cause of damage to several crops, and their potential role as model organisms for the study of biological, developmental and evolutionary studies, has granted species of holometabolous insects a darling place at the heart of the field of insect genomics. The following paragraphs describe a few highlights on the relevance and contributions of these genome-sequencing projects.

DIPTERA. Besides the extensively studied Drosophilid group, the mosquitoes have been the only other group within the order subjected to genome-sequencing projects. I will leave the highlights of extensive contributions the research in Drosophilid species has made to science, a subject to be further explored by the interested reader on the many, many pages available. The range and scope of knowledge obtained from research on this group has been discussed on a myriad of manuscripts including documents in many scientific journals (including articles by FlyBase Consortium 2002, Celniker and Rubin 2003, Ashburner and Bergman 2005, and Drysdale and FlyBase Consortium, 2008) and even a book that goes as far as describing the conflicting personalities of the researchers involved in the sequencing of the Drosophila Genome (Ashburner, 2006), among many others. As noted by Ashburner and Bergman (2005), one of the most important contributions of these projects was the establishment of unprecedented methodologies such as whole-genome shotgun sequencing and genome annotation by a community "jamboree", which became the model to be followed by most genome projects after that. Beyond the *D. melanogaster* project, the genome sequences of other 23 species of Drosophila will provide more advances on sequencing strategies and techniques for comparative analyses in addition to a platform for advancing our understanding of critical biological processes and their evolutionary history.

Also from the infraorder *Muscomorpha*, although distantly diverged from Drosophila, the genomes of three other flies are currently being sequenced. *Cochliomyia homnivorax* is the new world screw-worm fly, a parasitic species that feeds on the living tissue of warm-blooded animals. The horn fly *Hematobia irritans* is a blood-feeding parasite of cattle. The former a predominant species of the tropics and the latter spread worldwide, both species represent a threat to the cattle industry throughout the globe and determining the genome sequences of these organisms will provide information necessary to develop new technologies for their control. The “third fly” is actually a group of them, the *Glossina* flies or tsetse flies. Flies of the genus *Glossina* are cyclical vectors of the pathogens that cause the human African trypanosomiasis (sleeping sickness) and nagana in animals. Because of the antigenic variation in the trypanosome no vaccines are available and they are unlikely to be developed (Askoy and IGGI, 2007). Hence determining the genome of the tsetse flies will contribute to our understanding of the fly-trypanosome interactions during establishment of infections in the fly and such genomic tools will become elemental in developing new vector-based approaches to combat these diseases. Last in this list but not of least importance, two species of phlebotomine sand flies, *Lutzomyia longipalpis* and *Phlebotomus papatasi*, have been selected for whole genome sequencing. Sand flies are the transmitting vectors of bacterial, viral and protozoan pathogens, which are the causing agents of emerging and re-emerging human diseases such as bartonellosis (which may lead to fatal anemia or Oroya fever), sand fly fevers and Leishmaniasis. Genomic analyses on these species will accelerate the progress of research on the interactions between vector, parasite and host. Kamhawi (2006) reported on the evolutionary adaptations that ensure the survival of the parasite that causes Leishmaniasis. Such adaptations include secretion of phosphoglycans, which protect the parasite from digestive enzymes, secretion of a neuropeptide that arrests midgut and hindgut peristalsis, and attaching to the midgut to avoid expulsion, among others. Uncovering the genome-sequences of these two species of sand flies will help with the identification of sand fly molecules pertinent to vector competence. Ivens and colleagues published the genome sequence of the parasite *Leishmania major* in 2005. Having the genome from both vector and parasite will enable us to take a global look at these interactions at a molecular level.

Culicinae and Anophelinae are the two most medically important known mosquito subfamilies. The first includes *Aedes aegypti,* which is responsible for the transmission of the arbovirus that causes yellow fever and dengue fever, and *Culex pipiens quinquefasciatus*, which is responsible for the transmission of the West Nile virus and the nematode responsible for lymphatic filariasis (*Wuchereria bancrofti*). *Anopheles gambiae* from the Anophelinae is the primary vector for transmission of malaria. Mosquito control is still the only viable strategy for preventing dengue and other mosquito-borne diseases. The draft genome sequence of *A. aegypti* represents a significant technical achievement, which will stimulate efforts to elucidate interactions at the molecular level between mosquitoes and the pathogens they transmit (Nene et al., 2007). Comparative analysis of all mosquito species will illuminate our understanding of mosquito chromosome evolution, gene and gene function identification specific to host-seeking and blood-feeding behavior, and innate immune response to pathogens encountered during blood-feeding behavior (Collins, 2008). As for the mosquito *A. gambiae*, outcomes of a genome-sequencing project combined with EST sequencing efforts will contribute to obtain substantial improvements on the control of malaria addressing three main aspects of mosquito biology: i) decreasing the number and longevity of infectious mosquitoes, ii) understanding what attracts them to human hosts and iii) decreasing the capacity of parasites to fully develop within them (Holt et al., 2002). This genome represents yet another valuable molecular entomology resource that will hopefully lead to an effective intervention in transmission of malaria and perhaps other mosquito-borne diseases. *Culex pipiens quinquefasciatus* is the vector of the nematode that causes lymphatic filariasis and of West Nile virus (encephalitis). Determining the genome sequence of this organism will also contribute to the identification of mosquito genes required for pathogen transmission and will facilitate the development of new strategies for combating and controlling these diseases. This sequence will complement the ones from the mosquitoes *A. gambiae* and *A. aegypti* providing essential information for phylogenetic inferences and comprehensive comparisons within the group. It will also help to improve our understanding of genes involved in their capacity to act as vectors of disease-causing viruses and their adaptation to insecticides through the development of resistance. The possibility of performing genomic comparative analyses with the sequences from the *C. p. quinquefasciatus*, *A.*

gambiae, *A. aegypti* and the hornfly (*H. irritans*) genomes will be of paramount importance to study the physiological adaptations of a hematophagous diet.

LEPIDOPTERA. The task of sequencing the genomes of species of the order Lepidoptera was the first attempt at generating information for comparative genomics and functional studies outside of the very well studied species of the order Diptera. Besides the displays of both beauty and evolutionary innovations offered by a large number of butterfly species, Lepidoptera includes insect pests of maize, cotton and soybean, among many other crops and garden plants. Examples of pest species are the helothines *Helicoverpa armigera, H. zea* and *Heliothis virensces* and the specialist *Manduca sexta*, which attacks species of the genus *Nicotiana* (tobacco). Lepidopteran species are considered models for a variety of biological processes such as interactions between plants and Lepidoptera, and between Lepidoptera and their pathogens (ILGP. 2002). This order also includes an array of species important for agricultural use such as silk harvesting of *Bombyx mori*, and for offering research opportunities to study different models of phenotypic variations such as wing patterns on the African butterfly *Bicyclus anynana* (Wijngaarden and Brakefield, 2000) and metapopulation ecology on the fragmented landscape inhabited by the Glanville fritillary butterfly *Melitaea cinxia* (Orsini et al., 2008). The genome sequence of *Bombyx mori* has been determined to coverage of 90 - 97% of all known silkworm genes (Mita et al., 2004 and Xia et al., 2004) and the two major repositories of data for this project are Silkbase (http://morus.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi) and the NCBI Silkworm Genome Project web page (http://www.ncbi.nlm.nih.gov/sites/entrez). In addition to *B. mori*, sequencing of the genomes of *Bicyclus anynana* and *Melitaea cinxia* is currently under way.

HYMENOPTERA AND COLEOPTERA. Uncovering the genome of the honey bee (HBGSC, 2006) opened up the possibility to study the species "from molecule to colony" (Wilson, 2006). Honey bee is a model species for social behavior and an essential species to global ecology as pollinators. The genome of *A. mellifera* was the third insect genome to be sequenced after Drosophila flies and the mosquito *Anopheles gambiae*. One of the most relevant quests for the honey bee research community, around which many research foci revolve, deals with the evolution of their social lifestyle. How did such lifestyle come into being? In fact, the reports from the HBGSC were our first genome-wide insights into the social life of insects. The main findings of research conducted using the first draft assembly of the *A. mellifera* genome were, among many others, a higher A+T and CpG contents than those of Drosophila and *A. gambiae* and the lack of major transposon families. The Consortium also reported on data suggesting that the genome of the honey bee evolves more slowly and seems to be more similar to vertebrates for circadian rhythm, RNAi and DNA methylation genes. There are fewer genes for innate immunity, detoxification enzymes, cuticle proteins and gustatory receptors, and more genes for odorant receptors than in all dipterans examined until 2006. Their findings also included novel genes for nectar and pollen utilization, as well as population genetics studies suggesting a novel African origin for the species. These features seem to be consistent with the ecology and social organization of the species. Conducting research that expands on the findings of an increase in odorant receptor genes and a decrease in gustatory receptor genes may play a role in understanding the honey bee's adaptation to the social life and the evolution of feeding behaviors as well as its interaction with the environment (Wilson, 2006). Beye and coworkers (2006) used the first draft of the honey bee genome sequence and the improved genetic maps to perform analyses on recombination rates, which led them to propose evolutionary explanations for an exceptionally high genome-wide recombination rate. Their analyses showed that the honey bee presents a genome-wide recombination rate of 19cM/Mb, and that it is approximately ten times higher than the one reported for humans, *D. melanogaster* and *C. elegans* (about 1.6 per chromosome pair); such recombination rate is not restricted to certain regions of the chromosomes or specific chromosomes. Even so, the relationship between GC content and recombination is consistent with that of mouse, human and fly, which may suggest a common cause or consequence of recombination. As mentioned before, one of the main points raised by the analyses on the honey bee genome sequence is that common types of transposons and retrotransposons are largely absent (HBGSC, 2006). Beye's team suggested that it is possible that the high recombination rate enables a more efficient removal of deleterious insertions, such as the ones caused by the insertion of mobile elements on functional genes. Lastly, this group reported that the genome-wide phenomenon of higher levels of recombination in the honey bee is not dependent on chromosome size, although it is associated with gene size. That is, introns tend to be larger in regions of low recombination. This may have developed as a mechanism to improve efficacy of selection in these regions (Beye et al., 2006). This is just one example of the many applications that have been made possible with the availability of the *Apis* genome.

The same is true for the possibilities already explored with the genome of the red flour beetle (*Tribolium castaneum*). *T. castaneum* belongs to the most specious and the most evolutionarily successful eukaryotic order, Coleoptera (Hunt et al., 2007). The genome sequence of this model beetle, and pest of every stored grain or dried food, was published in 2008 (TGSC). *Tribolium* became a model for the study of insect development thanks to many favorable traits. These traits include ease of culture, high fecundity, short life cycle, facility for genetic crosses and the findings that RNAi is systemically spread from the site of injection to neighboring tissues (Tomoyasu and Denell, 2004) and from females to their progeny (Bucher et al., 2002), which facilitates knockdown of specific genes. *Tribolium* embryos develop following a short-germ model in which additional segments are sequentially added from a posterior growth zone, and *Tribolium* larvae have eyes in a fully formed head and three pairs of thoracic legs (TGSC, 2008). This developmental plan and such mechanism of segmentation are different from what is seen in Drosophila, and are believed to be more representative of other insects and basal arthropods (Tautz, 2004). The recently sequenced genome has already provided the tools to explore fascinating opportunities within the *Tribolium* research community as well as other insect orders. Lorenzen and coworkers reported on their findings on research conducted on the maternal-effect, selfish genetic element *Medea1* and its association with a *Tc1* transposon using the *T. castaneum* genome as a point of reference. Twenty-three beetle strains from 15 countries worldwide were tested for *Medea1*-

associated maternal-lethal activity, and then subjected to sequence analysis in the vicinity of a *Tc1* insertion site. Sequence comparisons suggested that the current distribution of *Medea1* reflects global emanation after a single transpositional event in recent evolutionary time. The *Medea* system in *Tribolium* represents an unusual type of intragenomic conflict and could provide a useful vehicle for driving desirable genes into populations (Lorenzen et al., 2008). Its characterization was made possible thanks to the availability of genomic resources such as two BAC libraries (used to clone the region around *Medea1* locus through BAC walking) and the newly available genome, which was used as reference sequence to characterize the 21.5 Kb insertion. Wanner and Robertson (2008) on the other hand, utilized the sequence of the *T. castaneum* genome along with the sequences of the Drosophilid flies, the mosquitoes and the honey bee, to annotate a total of 65 gustatory receptor genes from the silkworm *Bombyx mori* genome. This review can only attempt to scratch the surface of the collection of analyses that have been made possible thanks to the availability of these two holometabolous genomes. Neither *Tribolium* nor *Apis* is a vector of human diseases; rather both play an important role on human agricultural practices with the former acting as a pest and the latter as a pollinator. The genome sequences of both species constitute a landmark on the advancement of insect genomics and of our understanding of the evolutionary forces that shaped the most successful group of metazoans on the face of the earth, the Insects.

Public Genome Projects on Species of Hemimetabolous Insects

One major concern when considering genome sequencing projects for hemimetabolous insect species is that approximately 70% of the genomes of species examined to date have very large genome sizes ranging from approximately 978 Mb for the termite *Hodotermes mossambicus*, up to over 16,000 Mb for a number of orthopterans, including the mountain grasshopper (*Podisma pedestris*) (Gregory et al., 2006, Gregory, 2006). Large genome sizes complicate assembly efforts due to increased amounts of non-coding sequence populated with repetitive areas. In addition to this, the prevalence of holometabolous species in insect genome-sequencing projects adds a level of complexity to the assemblies of hemimetabolous ones due to the nucleotide sequence divergence inherent to such long evolutionary distances. Computational techniques used to generate *ab initio* prediction gene sets with genomic data rely on sequence similarity estimates obtained through algorithmic comparisons of the queried genome against all putative proteins available for all other insect genomes. The absence of other hemimetabolous insect species may then present a challenge when trying to establish a confident gene set for these organisms.

Despite the potential challenges, four research groups have separately (but almost simultaneously) embarked in the task of sequencing the genomes of 4 hemimetabolous insects. These species include three hemipterans, the pea aphid, the Asian citrus psyllid and the blood-feeding vector of Chagas disease; the fourth hemimetabolous insect is the human body louse, from the order Phthiraptera. Their genome sizes are uncharacteristically smaller than that of the majority of hemimetabolous species studied to date, ranging from approximately 100 Mb to just under 700 Mb. These smaller genome sizes may offer a lesser challenge for assembly.

The pea aphid (*Acyrthosiphon pisum*) and the Asian citrus psyllid (*Diaphorina citri*) are the cause of crop damage representing enormous monetary losses each year. Aphids attack both common garden plants and several major crops, and are vectors for many plant viruses that cause more damage than the aphids themselves (Stern, 2008). *A. pisum* lives in obligate mutualism with the bacterium *Buchnera aphidicola*, a situation that may be similar to what might have happened in early organelle evolution, and the availability of a genome sequence will increase the reach of the studies on this relationship.

Johnston and colleagues (2007) estimated what is believed to be the smallest genomes of hemimetabolous insects. The genome sizes of human body louse (*Pediculus humanus humanus*) and head louse (*Pediculus humanus capitis*) were estimated at approximately 107 Mb using flow cytometry determinations. *P. humanus humanus* is the primary vector which transmits three very specialized and diverse bacteria, which cause the historical human diseases louse-borne relapsing fever, trench fever and epidemic typhus. Preliminary assemblies of genome sequences of the bacterial agents of epidemic typhus (*Rickettsia prowazekii*) and trench fever as well as endocarditis (*Borrelia quintana*), a common infection among the homeless are currently available; thus, determining the genome sequence of the body louse will contribute to the advancement of studies of host-vector-pathogen interactions. Additionally, comparing the genome sequences of the aphid, the body louse and *Rhodnius prolixus* will shed light on the relationship between hosts and their parasites and changes associated with smaller genome sizes, as well as will provide tools to identify conserved regions among species of the Paraneoptera (Paraneoptera, as defined by Gullan and Cranston, 2004).

Rhodnius prolixus is the triatomid bug responsible for transmitting the pathogen *Trypanosoma cruzi*, which is the cause of Chagas' disease (American Trypanosomiasis). The genome size of *R. prolixus* is approximately 670 Mb (Panzera et al., 2007). As is the case of research on the body louse, determining the genome of this hemipteran is a matter of medical as well as economical importance. This potentially fatal parasitic disease affects approximately 10 - 14 million or more people in Latin America, but is also a cause for epidemiologic concern for other countries such as the United States and Canada (Strosberg et al., 2007). Spelling out genomic information from these species may contribute to the development of new pharmaceuticals and research on the search for alternative ways of controlling the transmission of diseases. Comparative analysis between this, the genome sequence of the body louse (another blood-feeding species) and the genome sequences of the pea aphid and the Asian citrus psyllid, will also allow the study of mechanisms underlying cellular processes associated with feeding, digestion, excretion and reproduction in blood-feeding insects relative to those with phytophagus diets.

Determining the genome sequences of these four species will provide valuable insights into the biology of animal interactions with microbes, among many other aspects of their natural history. It will also provide an outgroup for the study of evolution of holometabolous insects and in combination with genome sequences from the latter, it will also allow identification of evolutionarily conserved genes, the study of evolutionary divergence and mechanisms of cell division and chromosome structure and function. Understanding the factors, both molecular and physiological, which allow insects to act as vectors may help in the development of novel biopharmaceuticals (Huebner et al., 2005). And the use of genomic data on studies on adaptation in the form of insecticide resistance might also help to address the need for alternative methods of pest control, either through development of improved and more specific pesticides or through the more effective use of other insect species as agents of biological control.

The large number of insect species, their biomass, diversity of adaptation, and ecological impact, support the structure and function of ecosystem and biodiveristy on the lands of the earth (Zhang et al., 2007). Through a computational method that uses genome analysis to characterize insect and eukaryote proteomes, Zhang and his team reported that stress and stimulus response proteins were found to constitute a higher fraction in the insect-specific ortholog than in the orthologs common to eukaryotes. They concluded that the prevalence of these types of proteins in the insect-specific sequences, plus a plethora of specific cuticle and pheromone/odorant binding proteins, might suggest that communication and adaptation to environments may distinguish insect evolution relative to other eukaryotes. However, the picture is not yet complete and the evolutionary steps that led the species of Insecta to be such a successful group, which has colonized such diverse environments, are not completely understood. As it has begun to surface, we hope that the wealth of knowledge that will be contributed by the field of insect genomics, the comparative analyses that will be possible once these sequences are all available along with the hypotheses that researchers in this field will be able to test, will very likely help us increase our understanding of these critical evolutionary events. The focus of section has been primarily to report on the progress of publicly available insect genome projects and to briefly review some of the advances in the field on Insect Genomics. More exhaustive reviews on the topic may be found in Heckel, 2003, Robertson, 2005 and Grimmelikhuijzen et al., 2007.

Comparative genomics and molecular evolution: Working hand in hand

The field of molecular evolution was originated from two very different disciplines: population genetics, and molecular biology; a combination of the theory of the evolutionary process and the empirical data to test such theories. Molecular evolution seeks to describe the dynamics of evolutionary change at the molecular level, the driving forces behind this process and the effects of the various molecular mechanisms on the evolution of genomes, genes and proteins (Graur and Li, 2000). Researchers in this field have studied the evolution of genes by measuring the rates of synonymous and non synonymous substitutions at each site, using the codon as the unit of evolutionary change. Four decades ago Kimura (1968) concluded that the only viable explanation for the reported very high values of the rate of evolutionary change, in terms of nucleotide substitutions, was that many of the mutations involved must have been neutral ones. In other words, most of the variation within and between species does not affect the fitness of the organisms (Nielsen, 2005). This led Kimura to propose later the neutral theory of evolution, which additionally stated that new mutations arising in the population may increase in frequency due to random factors. This process is called genetic drift (Kimura, 1983). Since then, much debate has existed about the importance of natural selection in molecular evolution, and several methods have been designed to test for deviations from the neutral theory in an attempt to detect molecular adaptation (reviewed in Yang and Bielawski, 2000, Nielsen, 2001 and Nielsen, 2005) and re-examine the neutral theory of molecular evolution. How much of the sequence variation reported among species -even closely related ones- is the result of molecular adaptation, and how much is it the result of genetic drift? The publication of many genome sequences and the increasingly large amounts of DNA sequence diversity data available may bring researchers a step closer to answering the questions raised by this "neutralist-selectionist" debate (Eyre-Walker, 2006).

Gene duplication is widely accepted to be a major contributor to the origin of evolutionary novelties (Ohno, 1970). Many research groups have embarked on the task of exploring how duplicated genes survive acquiring novel functions, or become part of what Ohno believed was a collection of "now-extinct genes" making up the majority of our genomes (e.g.: Force et al., 1999, Doyle and Gaut, 2000, Lynch and Force, 2000, Lynch et al., 2001). These studies have demonstrated that a pair of gene duplicates may succumb to one of four evolutionary fates. One possibility is that i) both copies of the newly duplicated gene will be kept, and more RNA or more protein products constitute a benefit for the organism's fitness. In other instances while one copy maintains the original function ii) the other copy may be silenced by losing its function, iii) or it may adopt a new beneficial function. Another possibility is that iv) both copies may be kept through the fixation of complementary loss-of-function alleles, in which case the original function of the gene may be divided between the two in a process that Lynch and Force called subfunctionalization.

Duplicate genes are often referred as paralogous genes that form gene families (Zhang, 2003), which vary in the number of members across all taxa. For example, on the one hand the *Deleted in AZoospermia* (*DAZ*) gene family consists of two members, *DAZ* and *DAZL1*. Azoospermia is believed to be the most common form of infertility in human males (Shinka and Nakahori, 1996). *DAZ* is located in the Y chromosome and is found only in Old World Monkeys; *DAZL1*, is an autosomal gene found in all vertebrates. On the other hand the family of odorant receptor proteins in the honey bee (*A. mellifera*) is made of 170 genes (seven of which are pseduogenes), most of which are localized in tandem arrays and are divided onto 5 subfamilies. Such expansion of odorant receptors compared to the 60-80 receptors found in species of Diptera lies at the core of their remarkable olfactory abilities such as perception of several pheromone blends, kin recognition signals, and diverse floral odors. (Robertson and Wanner, 2006). Most animals are found to share several families of genes that regulate major aspects of body pattern (Carroll, 2000). One example is the family of homeotic genes that make up the *Hox* complex. Studies on these families of transcription factors and their signaling pathways and on the changes in number, regulation, or function of members of these families over the course of evolution have facilitated the study of the genetic basis of animal diversity (Carroll, 2000).

One hundred and eighty nucleotides constituting the homeobox, code for a highly conserved amino acid DNA-binding motif, the homeodomain, found in a family of related genes that encode transcription factors (Gehring, 1987). The *Hox* genes are a subset of these homeodomain-encoding transcription factors, which determine cell fate during the development of the animal embryo. Each of these genes is expressed along the anterior-posterior axis of the developing embryo to collectively determine segment identity in distinct domains by an orchestrated and complex interaction. With a few exceptions, these domains generally correspond spatially and temporally with their localization on the chromosome (Kaufman, 1990). A comprehensive review on the overall expression patterns of *Hox* genes in arthropods can be found Hughes and Kaufman, 2002. It is widely accepted that all paralogs of the *Hox* family of genes originated from a single ancestral gene by duplication. All arthropods and the Onychophora (velvet worms) share almost identical sets of *Hox* genes and most protostomes and deuterostomes (except vertebrates) posses roughly equivalent clusters of *Hox* genes possibly dating back to at least their Precambrian billaterial ancestor (Carroll, 2000). It is hypothesized that two rounds of genome-duplication events occurred early in the evolution of chordates and vertebrates (Hoegg and Meyer, 2005), turning a single ancestral cluster into multiple clusters ranging from four in tetrapods up to eight in ray-finned fish (Amores et al., 2004) and about fourteen in teraploid salmonid species (Moghadam et al., 2005). The ancestral arthropod had a single cluster of ten paralogs of the *Hox* family (Akam et al., 1994). In the grasshopper *Schistocerca gregaria* (Ferrier and Akam, 1996), the red flour beetle *T. castaneum* (Brown et al., 2002), the honey bee

A. mellifera (Dearden et al., 2006), and the jewel wasp *N. vitripennis* (Munoz-Torres, Unpublished data) *Hox* genes remain organized in a single cluster, and at least in the beetle and in both hymenopterans it is known that all members of the cluster are transcribed in the same orientation from the leading (positive) DNA strand. Insect evolutionary history shows that this cluster was split in the dipteran lineage (Lewis et al., 1980a, 1980b, Lewis, 1978) and in this group genes are transcribed in both orientations. A duplication event of a single gene (*Hox3*) and further loss of *Hox*-like function took place giving rise to the *zerknullt* (*zen*) gene, of which there are two copies in Drosophila (Falciani et al., 1996) and in *Tribolium* (Brown et al., 2002). *Zen* genes are involved in dorsoventral patterning in Cyclorrhaphan flies. *Bicoid* (*bcd*) encodes a morphogen for the anterior development of the Drosophila embryo, and is a maternal gene believed to have also evolved from the *Hox3* gene (Stauber et al., 1999). Another homeobox gene found throughout all arthropod lineages, but which has also lost its function as a *Hox* gene, is *fushi-tarazu* (*ftz*). *Ftz* acts as a Pair-Rule segmentation gene in Drosophila (Telford, 2000).

Over evolutionary time-scales, changes in the function or expression of these genes are associated with the diversification of segmental structures along the animal anterior-posterior axis (Hersh, 2007). Comparative studies of *Hox* genes have yielded much information regarding the genetic changes that lie behind the evolution of the arthropod body plan (Averof, 2002, Hughes and Kaufman, 2002, Deutsch and Mouchel-Viehl, 2003, Hughes et al., 2004, Angelini and Kaufman, 2005). In arthropods, variations on the expression patterns of *Hox* genes and on the regulation of their downstream targets govern the differences along the evolution of segmental specialization (Averof, 2002). These changes include large shifts in their regional domains of expression and the evolution of finer differences in their expression within individual segments. Despite the detailed picture available for the current state of *Hox* clusters, much controversy still exists as to the evolution of these clustered arrangements (Hersh, 2007). Contrasting models suggest on the one hand that the highly organized, compact clusters of vertebrates are derived from ancestral clusters that were less compact and less well organized (Duboule, 2007); on the other hand, evolution of *Hox* clusters is explained with the unequal crossing over that expanded a simple cluster consisting of only *Hox1* and *Hox 9* genes into more gene-rich clusters; this latter model also assumes that the original clusters must have been organized, not disorganized, based on conservation of spatial colinearity (W. Gehring as reported in Hersh, 2007). At any rate, what happened to the newly formed copies after the duplication events? Perhaps events of subfunctionalization following gene duplications led the *Hox* genes to their current patterns of expression. It is possible that following the duplication events, there was relaxation of the constraining forces imposed by these genes' roles in embryonic development, which might have led to a change of roles for some of the genes in the cluster. Perhaps also expansion and contraction of various genomic sequences may be just as important a mechanism of phenotypic evolution as changes at the nucleotide level. Spanning a genomic region of 1.68 Mb (Munoz-Torres, Unpublished data), the *Hox* cluster in *N. vitripennis* is slightly larger than that of the honey bee (1.37 Mb as reported by Dearden et al., 2006), and both are much larger than all other holometabolus insects studied to date. Much of these differences are accounted for by expansions of intronic and intergenic regions in the Hymenoptera, compared to all other insects studied to date. It is possible that besides the nucleotide changes these genome expansions also had a role in solidifying the separation of the Hymenoptera from all other lineages since they last shared a common ancestor. Given the more basal position of the Hymenoptera with regards to the orders Coleoptera, Lepidoptera and Diptera (as discussed above) this may indicate that a series of contractions of this genomic region in the latter after the Hymenoptera separated from their common ancestor may have occurred. Analyses of these non-coding regions may shed light on the role of *cis*-regulatory elements they contain in the evolution of morphological changes, as suggested by Averof (2002). It is possible that differences in the size of intronic and intergenic regions also have an effect on *cis*-regulatory elements and the way these interact with the genes they regulate, in this case a single *Hox* gene. These differences may have not represented changes in fitness (in other words, may have been neutral) during early stages of the duplication events if such regulatory regions were still interacting with their target genes. But if these regulatory elements became too dispersed or shuffled around enough to interrupt their role as regulator of target genes as a result of further duplication events, it may have had an important role on the evolution of gene functions in the long run. Studies on the origin and evolution of genes either by scrutinizing on the expansions of the single cluster in arthropods, or by looking at the molecular evolutionary history shaping the organization and function of individual genes from the cluster are now possible thanks to the availability of genomic data (Wolfe and Li, 2003, Averof 2002). Measurements of the rate of evolutionary changes between all

Hox paralogs in one species or between orthologs of the same *Hox* among species, at the nucleotide level, may yield information on the nature of the selective pressure acting upon the cluster. This may ultimately contribute to better understand the evolutionary forces that shaped these arrangements and provide a better look at the big picture in how these changes shaped the diversity we see today in the arthropod body plan.

References

Akam, M, M Averof, J Castelli-Gair, R Dawes, F Falciani, and D Ferrier. 1994. The evolving role of Hox genes in arthropods. *Dev Suppl* 209-215.

Amores, A, A Force, YL Yan, L Joly, C Amemiya, A Fritz, RK Ho, J Langeland, V Prince, YL Wang, M Westerfield, M Ekker, JH Postlethwait. 1998. Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282:**1711-1714.

Angelini, DR and TC Kaufman. 2005.Comparative Developmental Genetics and the Evolution of Arthropod Body Plans. *Annu Rev Genet* **39**:95-119.

Ashburner, M. 2006. Won for All: How the Drosophila Genome Was Sequenced. Cold Spring Harbor Laboratory Press. Woodbury, NY. USA. 100p.

Ashburner, M and CM Bergman. 2005. *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. *Genome Res* **15**(12):1661-1667.

Askoy, S, N Hall, and M Berriman on behalf of The International *Glossina* Genomics Initiative (IGGI) Community. 2007. A proposal for tsetse fly genome projects. White paper. 8p.

Averof, M. 2002. Arthropod *Hox* genes: insights on the evolutionary forces that shape gene functions. *Curr Op Genet Dev* **12**:386-392.

Beye, M, I Gattermeier, M Hasselmann, T Gempe, M Schioett, JF Baines, D Schlipalius, F Mougel, C Emore, O Rueppell, A Sirviö, E Guzmán-Novoa, G Hunt, M Solignac, RE Page Jr. 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Res* **16**(11):1339-44.

Boysen, C, MI Simon and L Hood. 1997. Analysis of the 1.1-Mb Human α/δ T-Cell Receptor Locus with Bacterial Artificial Chromosome Clones. *Genome Res* **7**:330-338.

Brown, S, JP Fellersb, TD Shippy, EA Richardson, M Maxwell, JJ Stuart, and RE Denella. 2002. Sequence of the *Tribolium castaneum* Homeotic Complex: The Region Corresponding to the *Drosophila melanogaster Antennapedia* Complex. *Genetics* **160**:1067-1074.

Bucher, G, J Scholten and M Klingler. 2002. Parental RNAi in *Tribolium* (Coleoptera). *CB* **12**:R85-R86.

Budiman, MA, L Mao, TC Wood and RA Wing. 2000. A Deep-Coverage Tomato BAC Library and Prospects Toward Development of an STC Framework for Genome Sequencing. *Genome Res* **10**:129-136.

Cai, L, JF Taylor, DS Gallagher, SS Woo and SK Davis. 1995. Construction and Characterization of a Bovine Bacterial Artificial Chromosome Library. *Genomics* **29**:413-425.

Campbell, CL, WC Black, AM Hess and BD Foy. 2008. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* **9**:425-441.

Carroll, SB. 2000. Endless Forms: The Evolution of Gene Regulation and Morphological Diversity. *Cell* **101**:577-580.

Celniker, SE and GM Rubin. 2003. The *Drosophila melanogaster* Genome. *Ann Rev of Genom Human Genet* **4**:89-117.

Chew, JS, C Oliveira, JM Wright and MJ Dobson. 2002. Molecular and cytogenetic analysis of the telomeric (TTAGGG)n repetitive sequences in the Nile tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). *Chromosoma* **111**:45-52.

Collins, FH. 2004. *Culex pipiens quinquefasciatus* genome project. Center for Global Health and Infectious diseases, University of Notre Dame. White paper. 10p.

Collins, FH. 2008. The Broad Institute *Aedes aegypti* Genome Project Information. http://www.broad.mit.edu/annotation/genome/aedes_aegypti.2/Info.html.

Dearden, PK, MJ Wilson, L Sablan, PW Osborne, M Havler, E McNaughton, K Kimura , NV Milshina, M Hasselmann, T Gempe, M Schioett, SJ Brown, CG Elsik, PW Holland, T Kadowaki, M Beye. 2006 Patterns of conservation and change in honey bee developmental genes. *Genome Res* **16**:1376-1384.

Deutsch, JS, E Mouchel-Vielh. 2003. Hox genes and the crustacean body plan. *Bioessays* **25**(9):878-87.

Duboule, D. 2007. The rise and fall of *Hox* gene clusters. *Development* **134**:2549-2560.

Drysdale, R and FlyBase Consortium. 2008. FlyBase: a database for the Drosophila research community. *Methods Mol Biol* **420**:45-59.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol* **21**(10):569-575.

Falciani, F, B Hausdorf, R Schröder, M Akam, D Tautz, R Denell, S Brown. 1996. Class 3 Hox genes in insects and the origin of zen. *Proc Natl Acad Sci USA* **93**:8479-8484.

Ferrier, DEK and M Akam. 1996. Organization of the *Hox* gene cluster in the grasshopper, *Schistocerca gregaria*. *Proc Natl Acad Sci USA* **93**:13024-13029.

Ferrier, D and C Minguillon. 2003. Evolution of the Hox/ParaHox gene clusters *Int J Dev Biol* **47**:605-611.

FlyBase Consortium. 2002. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30**(1):106-108.

Force, A, M Lynch, FB Pickett, A Amores, YL Yan, J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4):1531-1545.

Gehring, WJ. 1987. Homeo boxes in the study of development. *Science* **236**:1245-1252.

Graur, D, and Li W-H. 2000. Fundamentals of molecular evolution. *Sinauer Associates*, Sunderland, MA, USA. 481p

Gregory, TR, JA Nicol, H Tamm, B Kullman, K Kullman, IJ Leitch, BG Murray, DF Kapraun, J Greilhuber, and MD Bennett. 2006. Eukaryotic genome size databases. *Nucleic Acids Res* **35**(Database issue):D332-338.

Gregory, TR. 2006. Animal Genome Size Database. http://www.genomesize.com.

Grimaldi, DA and MS Engel. 2005. Evolution of the Insects. Cambridge University Press. New York, NY. USA. 755p.

Grimmelikhuijzen, CJP, G Cazzamali, M Williamson and F Hauser 2007. The promise of insect genomics. *Pest Manag Sci* **63**:413-416.

Gullan, PJ and P Cranston. 2004. The Insects. An outline of Entomology. 3rd Ed. Blackwell Pub. Malden, MA. USA. 505p.

Heckel D. 2003. Genomics in Pure and Applied Entomology. *Annu Rev Entomol* **48**:235-260.

Hersh, B. 2007. *Hox* en Provence. *Dev Cell* **13**:763-768.

Hoegg, S, A Meyer. 2005. *Hox* clusters as models for vertebrate genome evolution. *Trends Genet* **21**(8):421-424.

Holt, RA, GM Subramanian, A Halpern, GG Sutton, R Charlab, DR Nusskern, P Wincker, AG Clark, JMC Ribeiro, R Wides, S L. Salzberg, B Loftus, M Yandell, WH Majoros, DB Rusch, Z Lai, CL Kraft, JF Abril, V Anthouard, P Arensburger, PW Atkinson, H Baden, V de Berardinis, D Baldwin, V Benes, J Biedler, C Blass, R Bolanos, D Boscus, M Barnstead, S Cai, A Center, K Chaturverdi, GK Christophides, MA Chrystal, M Clamp, A Cravchik, V Curwen, A Dana, A Delcher, I Dew, CA Evans, M Flanigan, A Grundschober-Freimoser, L Friedli, Z Gu, P Guan, R Guigo, ME Hillenmeyer, SL Hladun, JR Hogan, YS Hong, J Hoover, O Jaillon, Z Ke, C Kodira, E Kokoza, A Koutsos, I Letunic, A Levitsky, Y Liang, JJ Lin, NF Lobo, JR Lopez, JA Malek, TC McIntosh, S Meister, J Miller, C Mobarry, E Mongin, SD Murphy, DA O'Brochta, C Pfannkoch, R Qi, MA Regier, K Remington, H Shao, MV Sharakhova, CD Sitter, J Shetty, TJ Smith, R Strong, J Sun, D Thomasova, LQ Ton, P Topalis, Z Tu, MF Unger, B Walenz, A Wang, J Wang, M Wang, X Wang, KJ Woodford, JR Wortman, M Wu, A Yao, EM Zdobnov, H Zhang, Q Zhao, S Zhao, SC Zhu, I Zhimulev, M Coluzzi, A della Torre, CW Roth, C Louis, F Kalush, RJ Mural, EW Myers, MD Adams, HO Smith, S Broder, MJ Gardner, CM Fraser, E Birney, P Bork, PT Brey, JC Venter, J Weissenbach, FC Kafatos, FH Collins, SL Hoffman. 2002. The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. *Science* **298**(5591):129-149.

The Honey Bee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera. Nature* 443:931-949.

Hong, YS, JR Hogan, X Wang, A Sarkar, C Sim, BJ Loftus, C Ren, ER Huff, JL Carlile, K Black, HB Zhang, MJ Gardner and FH Collins. 2003. Construction of a BAC library and generation of BAC end sequences-tagged connectors for genome sequencing of the African malaria mosquito *Anopheles gambiae*. *Mol Gen Genomics* **268**:720-728.

Huebner, E and The *Rhodnius* Research Community. 2005. The case for sequencing the genome of the blood-feeding hemipteran insect, *Rhodnius prolixus*. White paper. 14p.

Hughes, CL, PZ Liu, TC Kaufman. 2004. Expression patterns of the rogue *Hox* genes *Hox3*/*zen* and *fushi tarazu* in the apterygote insect *Thermobia domestica*. *Evol Dev* **6**:393-401.

Hughes, CL and TC Kaufman. 2002. Hox genes and the evolution of the arthropod body plan. *Evol Dev* **4**(6):459-499.

Hunt, T, J Bergsten, Z Levkanicova, A Papadopoulou, O St. John, R Wild, P M Hammond, D Ahrens, M Balke, MS Caterino, J Gómez-Zurita, I Ribera, TG Barraclough, M Bocakova, L Bocak, AP Vogler. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* **318**:1913-1916.

International Lepidopteran Genome Project. 2002. Why Lepidoptera? The "*Bombyx* +" project. White paper. 13p.

Ivens, AC, CS Peacock, EA Worthey, L Murphy, G Aggarwal, M Berriman, E Sisk, M-A Rajandream, E Adlem, R Aert, A Anupama, Z Apostolou, P Attipoe, N Bason, C Bauser, A Beck, SM Beverley, G Bianchettin, K Borzym, G Bothe, CV Bruschi, M Collins, E Cadag, L Ciarloni, C Clayton, RM Coulson, A Cronin, AK Cruz, RM Davies, J De Gaudenzi, DE Dobson, A Duesterhoeft, G Fazelina, N Fosker, AC Frasch, A Fraser, M Fuchs, C Gabel, A Goble, A Goffeau, D Harris, C Hertz-Fowler, H Hilbert, D Horn, Y Huang, S Klages, A Knights, M Kube, N Larke, L Litvin, A Lord, T Louie, M Marra, D Masuy, K Matthews, S Michaeli, JC Mottram, S Müller-Auer, H Munden, S Nelson, H Norbertczak, K Oliver, S O'neil, M Pentony, TM Pohl, C Price, B Purnelle, MA Quail, E Rabbinowitsch, R Reinhardt, M Rieger, J Rinta, J Robben, L Robertson, JC Ruiz, S Rutter, D Saunders, M Schäfer, J Schein, DC Schwartz, K Seeger, A Seyler, S Sharp, H Shin, D Sivam, R Squares, S Squares, V Tosato, C Vogt, G Volckaert, R Wambutt, T Warren, H Wedler, J Woodward, S Zhou, W Zimmermann, DF Smith, JM Blackwell, KD Stuart, B Barrell, PJ Myler. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**:436-442.

Jiménez, LV, BK Kang, B de Bruyn, DD Lovin and DW Severson. 2004. Characterization of an *Aedes aegypti* bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence Plasmodium susceptibility. *Insect Mol Biol* **13**(1):37-44.

Johnston, JS, KS Yoon, JP Strycharz, BR Pittendrigh, CJ Marshall. 2007. Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any Hemimetabolous insect reported to date. *J Med Entomol* **44**(6):1009-1012.

Kaufman, TC, MA Seeger and G Olsen. 1990. Molecular and genetic organization of the *Antennapedia* gene complex of *Drosophila melanogaster*. *Adv Gen* **27**:309-362.

Kamhawi, S. 2006. Phlebotomine sand flies and *Leishmania* parasites: friends or foes? *Trends in Parasitology* **22**(9):439-445.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**:624-626.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University, New York, NY, USA. Press. 367p.

Koike, Y, K Mita, MG Suzuki, S Maeda, H Abe, K Osoegawa, PJ deJong, T Shimada. 2003. Genomic sequence of a 320-kb segment of the Z chromosome of *Bombyx mori* containing a kettin ortholog. *Mol Genet Genomics* **269**(1):137-49.

Kuska, B. 1998. Beer, Bethesda, and biology: how "genomics" came into being. *J Natl Cancer Inst* **90**:93.

Lewis, EB.1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**:565-570.

Lewis, RA, TC Kaufman, RE Denell, P Tallerico.1980a. Genetic Analysis of the *Antennapedia* Gene Complex (Ant-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. I. Polytene Chromosome Segments 84b-D *Genetics* **95**(2):367-381.

Lewis, RA, BT Wakimoto, RE Denell, TC Kaufman. 1980b. Genetic Analysis of the *Antennapedia* Gene Complex (Ant-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. II. Polytene Chromosome Segments 84A-84B1, 2. *Genetics* **95**(2):383-397.

Lorenzen, MD, A Gnirke, J Margolis, J Garnes, M Campbell, JJ Stuart, R Aggarwal, S Richards, Y Parki and RW Beeman. 2008. The maternal-effect, selfish genetic element *Medea* is associated with a composite *Tc1* transposon. *Proc Natl Acad Sci USA* **105**(29):10085-10089.

Lucigen Corporation. 2008. Random Shear BAC Library Construction. White Paper. 7p.

Lynch, M and A Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**(1):459-73.

Lynch, M, M O'Hely, B Walsh, A Force. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**(4):1789-804.

Mahtani, MM and Willard HF. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Research* **8**:100-110.

Mita, K, M Kasahara, S Sasaki, Y Nagayasu, T Yamada, H Kanamori, N Namiki, M Kitagawa, H Yamashita, Y Yasukochi, K Kadono-Okuda, K Yamamoto, M Ajimura, G Ravikumar, M Shimomura, Y Nagamura, T Shin-I, H Abe, T Shimada, S Morishita, T Sasaki. 2004. The genome sequence of silkworm, *Bombyx mori*. *DNA Res* **11**(1):27-35.

Mita, K, M Mitsuoki, O Kazuhiro, K Yoshiko, N Junko, MG Suzuki and T Shimada. 2002. Construction of an EST database for *Bombyx mori* and its applications. *Current Science* **83**(4):426-431.

Moghadam, HK, MM Ferguson and RG Danzmann. 2005. Evidence for Hox Gene Duplication in Rainbow Trout (*Oncorhynchus mykiss*): A Tetraploid Model Species. *J Mol Evol* **61**(6):804-818.

Nene, V, JR Wortman, D Lawson, B Haas, C Kodira, Z (Jake) Tu, B Loftus, Z Xi, K Megy, M Grabherr, Q Ren, EM Zdobnov, NF Lobo, KS Campbell, SE Brown, MF Bonaldo, J Zhu, SP Sinkins, DG Hogenkamp, P Amedeo, P Arensburger, PW Atkinson, S Bidwell, J Biedler, E Birney, RV Bruggner, J Costas, MR Coy, J Crabtree, M Crawford, B Debruyn, D Decaprio, K Eiglmeier, E Eisenstadt, H El-Dorry, WM Gelbart, SL Gomes, M Hammond, LI Hannick, JR Hogan, MH Holmes, D Jaffe, JS Johnston, RC Kennedy, H Koo, S Kravitz, EV Kriventseva, D Kulp, K Labutti, E Lee, S Li, DD Lovin, C Mao, E Mauceli, CF Menck, JR Miller, P Montgomery, A Mori, AL Nascimento, HF Naveira, C Nusbaum, S O'leary, J Orvis, M Pertea, H Quesneville, KR Reidenbach, YH Rogers, CW Roth, JR Schneider, M Schatz, M Shumway, M Stanke, EO Stinson, JM Tubio, JP Vanzee, S Verjovski-Almeida, D Werner, O White, S Wyder, Q Zeng, Q Zhao, Y Zhao, CA Hill, AS Raikhel, MB Soares, DL Knudson, NH Lee, J Galagan, SL Salzberg, IT Paulsen, G Dimopoulos, FH Collins, B Birren, CM Fraser-Liggett, DW Severson. 2007. Genome Sequence of *Aedes aegypti*, a Major Arbovirus Vector. *Science* **316**(5832):1718-1723.

Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**(6):641-647.

Nielsen, R. 2005. Molecular Signatures of Natural Selection. *Annu Rev Genet* **39**:197-218.

Nobrega, MA and LA Pennacchio. 2003. Comparative genomics analysis as a tool for biological discovery. *J Physiol* **554**(1):31-39.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, New York, NY, USA. 160p.

Orsini, L, J Corander, A Alasentie and I Hanski. 2008. Genetic spatial structure in a butterfly metapopulation correlates better with past than present demographic structure. *Mol Ecol* **17**:2629-2642.

Osegawa, K, C-LShu, J Catanese, P de Jong. 2001. New procedures for construction of large insert BAC libraries. *Workshop*: Large-Insert DNA Libraries and Their Applications. IX International Plant and Animal Genome Conference. San Diego, CA.

Panzera, F, I Ferrandis, J Ramsey, PM Salazar-Schettino, M Cabrera, C Monroy, MD Bargues, S Mas-Coma, JE O'connor, VM Angulo, N Jaramillo and R Pérez. 2007. Genome size determination in chagas disease transmitting bugs (Hemiptera-Triatominae) by flow cytometry. *Am J Trop Med Hyg* **76**(3):516-521.

Robertson, HM. 2005. Insect genomes. *American Entomologist* **51:**166-171.

Robertson, HM and KW Wanner. 2006 The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res* **16**(11):1395-403.

Savard, J, D Tautz, S Richards, GM Weinstock, RA Gibbs, JH Werren, H Tettelin, MJ Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res* **16**(11):1334-1338.

Shinka, T, and Y Nakahori. 1996. The azoospermic factor on the Y chromosome. *Acta Paediatr Jpn* **38**:399-404.

Shizuya, H, B Birren, U J Kim, V Mancino, T Slepak, Y Tachiiri, and M Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA* **89**(18):8794-8797.

Stauber, M, H Jäckle, and U Schmidt-Ott. 1999. The anterior determinant *bicoid* of Drosophila is a derived *Hox* class 3 gene. *Proc Natl Acad Sci USA* **96**(7):3786-3789.

Stern, DL. 2008. Aphids. *CB* **18**(12):R504-R505.

Strosberg, AM, K Barrio, VH Stinger, J Tashker, JC Wilbur, L Wilson and K Woo. 2007. CHAGAS DISEASE: A Latin American Nemesis. Report prepared by the Institute for OneWorld Health for the Bill and Melinda Gates Foundation. San Francisco, CA. USA. 110p.

Tautz, D. 2004. Segmentation. *Dev Cell* **7**:301-312.

Telford, MJ. 2000. Evidence for the derivation of the Drosophila *fushi-tarazu* gene from a *Hox* gene orthologous to lophotrochozoan Lox5. *CB* **10**:349-352.

Tomkins, J, M Luo, G Fang, D Main, J Goicoechea, M Atkins, D Frisch, RE Page, E Guzman-Novoa, Y Yu, GJ Hunt and R Wing. 2002. New genomic resources for the honey bee (*Apis mellifera* L.): development of a deep-coverage BAC library and a preliminary STC database. *Genet Mol Res* **1**:306-316.

Tomoyasu, Y and RE Denell. 2004 Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev Genes Evol* **214**:575-578.

Tribolium Genome Sequencing Consortium. 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**(7190):949-955.

Wanner, KW and HM Robertson. 2008. The gustatory receptor family in the silkworm moth Bombyx mori is characterized by a large expansion of a single lineage of putative bitter receptors *Insect Mol Biol* **17**(6):621-629.

Wijngaarden, PJ and PM Brakefield. 2000. The genetic basis of eyespot size in the butterfly *Bicyclus anynana*: an analysis of line crosses. *Heredity* **85**:471-479.

Wilson, EO. 2006. How to make a social insect. *Nature* **443**:919-920.

Wolfe, KH and WH Li. 2003. Molecular evolution meets the genomics revolution. *Nat Genet* **33:**Suppl, 255-265.

Woo, SS, JM Jiang, BS Gill, AH Paterson, RA Wing. 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. *Nucleic Acids Res* **22**:4922-4931.

Xia, Q, Z Zhou, C Lu, D Cheng, F Dai, B Li, P Zhao, X Zha, T Cheng, C Chai, G Pan, J Xu, C Liu, Y Lin, J Qian, Y Hou, Z Wu, G Li, M Pan, C Li, Y Shen, X Lan, L Yuan, T Li, H Xu, G Yang, Y Wan, Y Zhu, M Yu, W Shen, D Wu, Z Xiang, J Yu, J Wang, R Li, J Shi, H Li, G Li, J Su, X Wang, G Li, Z Zhang, Q Wu, J Li, Q Zhang, N Wei, J Xu, H Sun, L Dong, D Liu, S Zhao, X Zhao, Q Meng, F Lan, Huang X, Li Y, Fang L, Li C, Li D, Sun Y, Zhang Z, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Li J, Ye J, Chen H, Zhou Y, Liu B, Wang J, Ye J, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Li S, Wang J, Wong GK, Yang H and the Biology Analysis Group. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**(5703):1937-1940.

Yang, ZH and JP Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**(12):496-503.

Yu, Y, JP Tomkins, R Waugh, DA Frisch, D Kudrna, A Kleinhofs, RS Brueggeman, GJ Muehlbauer, RP Wise, RA Wing. 2000. A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theor Appl Genet* **101**:1093-1099.

Yuan G, C Qi, P Junsong, L Zheng, H Huanle, W Aizhong, S Rentao and C Run. 2008. Construction of a BAC library from cucumber (*Cucumis sativus* L.) and identification of linkage group specific clones. *Prog Nat Sci* **18**:143-147.

Zdobnov, EM and P Bork. 2006. Quantification of insect genome divergence. *TIG* **23**(1):16-20.

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**(6):292-298.

Zhang, G, H Wang, J Shi, X Wang, H Zheng, G Wong, T Clark, W Wang, J Wang, and Le Kang. 2007. Identification and characterization of insect-specific proteins by genome data analysis. *BMC Genomics* **8**(1):93.

Zhong, J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**(6):292-298.

Zhou, H, DK Thompson, Y Xu and JM Tiedje. 2004. Microbial Functional Genomics. John Wiley & Sons, Inc. Hoboken, NJ, USA. 590p.

CHAPTER TWO

CONSTRUCTION AND CHARACTERIZATION OF A BAC-LIBRARY FOR A KEY POLLINATOR, THE BUMBLEBEE *Bombus terrestris* L.

With kind permission from Springer Science+Business Media: *Insectes Sociaux*. Online First. October 11, 2008. Wilfert, L. [1, 5,6], Muñoz Torres, M.[2, 5], Reber-Funk, C.[1], Schmid-Hempel, R.[1, 7], Tomkins, J. [3], Gadau, J. [4], Schmid-Hempel P.[1], 2 Figures. DOI: 10.1007/s00040-008-1034-1. License number: 2071480913378. License Date: Nov 17, 2008.

Authors and Affiliations

Wilfert, L. [1, 5,6], Muñoz Torres, M.[2, 5], Reber-Funk, C.[1], Schmid-Hempel, R.[1, 7], Tomkins, J. [3], Gadau, J. [4], Schmid-Hempel P.[1]

[1] Institute of Integrative Biology (IBZ), ETH Zurich, CH-8092 Zurich, Switzerland; e-mail: lena.wilfert@ed.ac.uk

[2] Clemson University Genomics Institute, Clemson University, Clemson, SC 29634, USA

[3] Clemson Environmental Genomics Lab, Clemson University, Clemson, SC 29634, USA

[4] School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA

[5] These authors contributed equally to this work.

[6] current address: Institute of Evolutionary Biology, University of Edinburgh, UK

[7] contact for data repository of this BAC library

Abstract

The primitively social bumblebee *Bombus terrestris* is an ecological model species as well as an important agricultural pollinator. As part of the ongoing development of genomic resources for this model organism, we have constructed a publicly available bacterial artificial chromosome (BAC) library from males of a field-derived colony. We have shown that this library has a high coverage, which allows any particular sequence to be retrieved from at least one clone with a probability of 99.7%. We have further demonstrated the library's usefulness by successfully screening it with probes derived both from previously described *B. terrestris* genes and candidate genes from another bumblebee species and the honey bee. This library will facilitate genomic studies in *B. terrestris* and will allow for novel comparative studies in the social Hymenoptera.

Introduction

Many species of social insects are not only commercially important but have become model species for a wide range of basic questions such as social evolution, caste determination, sex ratio strategies, foraging behavior, social parasitism, ecological physiology, sensory ecology, evolutionary parasitology, ecological immunology, as well as pollination and community ecology. Bumblebees have served as subjects of study in all of these aspects (e.g. Goulson, 2003). Indeed, only the honey bee appears to rival the breadth of research done on this group. Despite the significant contributions that social insects have made, the genomic resources – so common and well developed for many other groups of insects or vertebrates – are quite limited. With a few exceptions (Hoffman and Goodisman, 2007; Wang et al., 2007), important genomic resources (genome sequence, libraries of genomic DNA and expressed sequences) are available only for the honey bee. Here, we report on a substantial genomic resource for a social insect outside the honey bee, a high-coverage bacterial artificial chromosome (BAC) library of genomic DNA for the bumblebee *Bombus terrestris* L.

B. terrestris is a common European bumblebee species that is of importance as an ecological and evolutionary model organism (Goulson, 2003). The genomic tools promise to open up new avenues of research for ecological studies of organisms such as the bumblebee. For example, the evolution of sociality and caste determination is a research topic that has recently benefited from the application of genomic tools. Pereboom et al. (Pereboom et al., 2005) have identified genes that are differentially expressed in the development of queens and workers. This allows for comparative studies of caste determination on a detailed molecular level. Much recent research on bumblebees has focused on ecological immunity (Sadd and Schmid-Hempel, 2006) and host-parasite interactions (i.e. Baer and Schmid-Hempel, 2001; Ruiz-Gonzalez and Brown, 2006). This research has been complemented by quantitative genetic studies identifying quantitative trait loci (QTL) explaining a part of the phenotypic variation of fitness-relevant traits such as the strength of innate immune mechanisms (Wilfert et al., 2007b) and susceptibility to a protozoan parasite (Wilfert et al., 2007a). Such tools raise the opportunity to study the maintenance of genetic variation for fitness-relevant traits involved in host-parasite interactions in natural populations (Schmid-Hempel, 2001). To comprehensively test such hypotheses by studying the molecular signature of evolution, we need to identify the genes underlying the quantitative variation in host resistance.

Identifying single genes in turn requires access to the physical genome. The genetic map used to identify QTLs, by contrast, is based on recombination distances. In *B. terrestris*, one centimorgan of genetic distance represents an average of 226 Kb of the physical genome (Wilfert et al., 2006). To integrate these two approaches, large-insert libraries such as bacterial artificial chromosome (BAC) libraries are extremely valuable tools because they allow the physical mapping of genes based on information from genetic linkage maps, expressed sequences or heterologous candidate genes. As previous studies have shown, this joint approach is very successful. For example, the gene underlying the sex determination locus has been isolated with the help of a BAC-library in the honey bee (Tomkins et al., 2002; Beye et al., 2003). In order to facilitate the identification and cloning of genes, and to eventually facilitate physical mapping and genome assembly, we have here constructed a high coverage BAC-library of *B. terrestris* as one of the first such tools outside of *Apis*.

Materials and Methods

High-molecular weight DNA from fresh haploid male pupae was prepared following a procedure adapted for honey bees (Tomkins et al., 2002). DNA was partially digested with the restriction enzyme HindIII. Size selection was performed on the fragments via two consecutive rounds of pulse field electrophoresis (PFGE). Fragments were then ligated into the vector pIndigoBAC 536 (Peterson et al., 2000). Vectors were transformed into E. coli DH10B cells using electroporation. Recombinant colonies were picked using a Genetix Q-bot and stored individually in 384 well plates at -80° C. Additionally high-density colony filters for hybridization-based screening of the BAC library were prepared using a Genetix Q-Bot. Clones were arrayed in double spots using a 4 x 4 array with 6 fields, on 11.5 x 22.5 cm – Hybond N+ filters (Amersham). This pattern allows 18,432 clones to be represented per filter. Colony filters were grown and processed using standard techniques for alkaline lysis (Sambrook et al., 1989).

Library Screening

To screen the library, we developed PCR products from three *B. terrestris*-specific sequences (*Arginine Kinase, Elongation Factor 1 Alpha* and *Longwave-Rhodopsin*), one from a sequence from *B. ignitus* (*Defensin*) and two from *Apis mellifera* genes (*Relish* and *Dscam*) and used them as probes for hybridization. Primers and accession numbers of these genes are detailed in Table 2.1. The PCR products obtained were 200 – 400 bp in length; approximately 100 ng of product were labeled independently using 30uCi of alpha-P32-dCTP following manufacturer's instructions with the DECAprime II random priming DNA labeling kit (Ambion, Inc. ABI, Foster City, CA, USA). Hybridization of colony filters was performed using standard techniques (Sambrook et al., 1989) with the following modifications: hybridizations were performed for at least 16 h at 60º C (with *B. terrestris* probes) and 55º C (for heterologous probes); filters were washed twice at corresponding temperatures (30 min per wash) in a 2X SSC/0.1% SDS solution first, and in a 1X SSC/0.1% SDS solution the second time. Hybridized BAC-filters were imaged in a Storm Scanner (GE Healthcare, Piscataway, NJ, USA) and positive hits were scored with HybSweeper (Lazo et al., 2005).

BAC-DNA preparation and fingerprinting analyses.

Positive clones were fingerprinted using techniques established by Chen et al. (Chen et al., 2002) and Marra et al. (Marra et al., 1997). Briefly, DNA from BAC clones was prepared from 900 µL cultures of Terrific Broth (GIBCO)-Chloramphenicol (12.5 ug/uL) in 96-well format, inoculated with 1.5 µL of BAC freezer stocks. After 18 h, cultures were treated with a modified alkaline lysis method. Samples for fingerprinting were digested with the restriction endonuclease HindIII, electrophoresed on 1% agarose gels for 15 h at 60 V, and stained with Sybr Gold (Invitrogen) for 1 h. Gels were imaged in a Storm Scanner (GE Healthcare, Piscataway, NJ, USA). Fingerprinting data were scored using Image3 software (v.3.10, www.sanger.ac.uk/software/Image/).

To determine average insert size of the library, 192 clones were randomly selected from the library, DNA was digested using the endonuclease NotI (NEB, Ipswich, MA, US) and analyzed by PFGE.

Results

The bumblebee BAC library consists of 36’864 clones. The average insert size (n = 186) was 102.9 ± 28.5 Kb (see Figure 2.1) with a range of 40 – 220 Kb. PFGE analysis revealed that the library contains 3.1% empty clones (6 of 192 clones assayed for insert size. See Figure 2.2). The bumblebee genome has been estimated as being 625 Mb in size (Wilfert et al., 2006). The library thus has an expected coverage of 6x genome equivalents, allowing any one particular *B. terrestris* sequence to be recovered with a probability of 99.7% from at least one clone.

B. terrestris is a model organism for the evolution of the innate immune system. We therefore screened the BAC library with probes derived from candidate genes involved in antimicrobial defense pathways (NF-κB-like transcription factor *Relish*, *Defensin*) and in parasite recognition / phagocytosis (*Dscam*). As a positive control, we screened for genes that had previously been used to infer the phylogeny of the genus *Bombus* (Kawakita et al., 2004) (*Long-wave Rhodopsin, Arginine Kinase*, and *Elongation Factor 1 alpha*). An average of 27.5 ± 13.3 positive clones for each of six gene probes was retrieved, indicating a high redundancy of the library (Table 2.1).

The BAC clones identified by hybridization are expected to contain some false positives. To obtain the clones most likely to contain the genes of interest, positive BAC clones were fingerprinted with HindIII and then assembled into contigs using the Fingerprinted Contigs software (FPC (Soderlund et al., 2000)) at high stringency (using a tolerance value of 7 and a minimum cutoff value of 1e-10). This analysis allowed us to anchor genetic markers of the six genes analyzed here onto corresponding physical regions of the genome, represented by BAC clones. Markers generated from *Dscam, Defensin*, *Long-Wave Rhodopsin* and *Elongation Factor 1 alpha* were represented in four separate contigs containing 22, 8, 10 and 6 BAC clones respectively; the markers generated for *Arginine kinase* and *Relish identified* 9 and 8 clones respectively, which were represented in a single contig containing 17 BAC clones. These results provide supporting evidence for the quality of the library and constitute a first step in describing the location of these genes on the *B. terrestris* genome.

The FPC analysis allowed us to drastically reduce the number of candidate clones that will become targets for sequencing, i.e. from 52 clones identified in the hybridization screen to 9 clones for *Arginine Kinase*. A common cause of high numbers of false positives in hybridization screens is the use of degenerate primers, which leads to some degree of unspecific binding, to obtain clones containing candidate genes known only in related species. We have used this approach to identify candidate BAC clones for the isolation of immune genes in *B. terrestris* based on sequence information from the honey bee *A. mellifera*. Additionally, several of the genes we screened for - *Arginine Kinase*, *Defensin* and *Elongation Factor 1 alpha* - are known to be double copy genes in *A. mellifera* (Consortium, 2006) and *Nasonia vitripennis* (Genome Assembly 1.0, personal communication; Stephen Richards, Human Genome Sequencing Center, Baylor College of Medicine), this may also inflate the number of positive clones. FPC analysis is a powerful tool to deal with these issues common in genomic studies of non-model organisms: only those clones sharing statistically significant similar band patterns will become candidate clones for further analysis.

Based on the average insert size and the number of BAC clones generated, we have estimated a 6x coverage of the bumblebee genome. The mean number of positive clones after FPC analysis for the genes used in this screen indicates an average 10x coverage. This discrepancy can be explained by over- and underrepresentation of certain regions of the genome. Such biases in genome coverage include the use of a single enzyme, a partial restriction digestion of genomic DNA, and the unavailability of certain regions of genomes such as centromeres, highly repetitive sequences and telomeres, due to their lack of recognition sites for common restriction enzymes (Mahtani and Willard., 1998; Chew et al., 2002; Yuan et al., 2008). With an average 10x coverage of coding sequences, this library is likely to prove a valuable tool in the genomic analysis of *B. terrestris* and related social Hymenoptera.

Discussion

We here describe the construction and characterization of a BAC-library for the bumblebee *Bombus terrestris*. This high-quality library may serve as an important resource for genomic studies of bumblebees, such as gene isolation and genome mapping. We have screened the BAC library with several probes, demonstrating the library's usefulness as a genomic tool for *B. terrestris*. We could retrieve not only clones positive for already described sequences specific to described *B. terrestris* genes (*Long-wave Rhodopsin, Arginine Kinase,* and *Elongation Factor 1 alpha* (Kawakita et al., 2004)) but also for probes derived from another species from the genus Bombus (*Defensin* from *B. ignitus*) and from the distantly related honey bee *A. mellifera* (*Dscam* (Graveley et al., 2004) and *Relish*).

Comparative research into the social Hymenoptera stands to gain much by combining information from the sequenced genome of the honey bee and genomic information from related species (The Honey Bee Genome Consortium, 2006). Understanding the genetics of sex determination in the haplo-diploid Hymenoptera may prove to be a case in point. In the social Hymenoptera, sex is determined by a single complementary locus that triggers male development in hemizygous and homozygous embryos (Cook and Crozier, 1995), while other mechanisms are used in many families of the Hymenoptera (Heimpel and de Boer, 2008). In honey bees, the genetics of complementary sex determination (CSD) was first investigated using linkage mapping (Hunt and Page, 1994). To identify the responsible gene, the identified genetic region was fine-mapped (Hasselmann et al., 2001). With the help of a honey bee BAC library, the csd gene was then identified and demonstrated to be functional (Beye et al., 2003). Similarly, the sex determination locus in *B. terrestris* has been genetically mapped to an approximate location (Gadau et al., 2001). Using information from the honey bee and the BAC library we here describe, it will be possible to rapidly investigate the molecular and genetic nature of sex determination in bumblebees. This BAC library thus is not only a valuable tool for investigating the bumblebee genome, but vastly increases the potential for informative comparative studies in the social Hymenoptera.

B. terrestris BAC resources (library and high density filters) may be ordered from the Clemson University Genomics Institute (http://www.genome.clemson.edu/). The use of this BAC library should make reference to this paper. To maximize the information gained from this resource, a data repository for the BAC library is managed by R. Schmid-Hempel, ETH Zürich (rsh@env.ethz.ch).

Acknowledgements

The construction of the BAC library was financed by an ETH Research Grant (nr. TH-19/03-2 to PSH and JG).

Table 2.1. Hybridization results of six gene probes using the *Bombus terrestris* BAC library. Probes were generated by PCR. Positive clones were fingerprinted by HindIII and assembled into contigs.

Gene probe/ Accession Number	**Primers (5'-3')**	**Positive clones**	**Contig**	**Clones in contig**
Arginine Kinase (Kawakita et al., 2003) AF492888	F- GTTGACCAAGCYGTYTTGGA R- CATGGAAATAATACGRAGRTG	52	A	17
Defensin AY425599	F- GTGGCTCTTCTCTTTGTGGCTG R- CACTCTTCTTTGTCTATCGGCACG	21	B	8
Dscam (Graveley et al., 2004) AY686596	F- TTGGCTTTCACTTCTGGCGG R- TGCGGTCCACTTCCTTGATG	27	C	22
Elongation Factor 1 alpha (Kawakita et al., 2003) AF492955	F- GGACACAGAGATTTCATCAARAA R- TTGCAAAGCTTCRTGRTGCATTT	27	D	6
Long-wave Rhodopsin (Mardulyn and Cameron, 1999) AF091722	F- AATTGCTATTAYGARACNTGGGT R- ATATGGAGTCCANGCCATRAACCA	12	E	10
Relish XM_624623	F- TGGACGCTTTTCAGAATTGG R- GAGCTTCCAGAATGAGATATTCG	26	A	17

Figure 2.1: Histogram of insert size distribution of BAC clones (n = 186) of the bumblebee BAC-library.

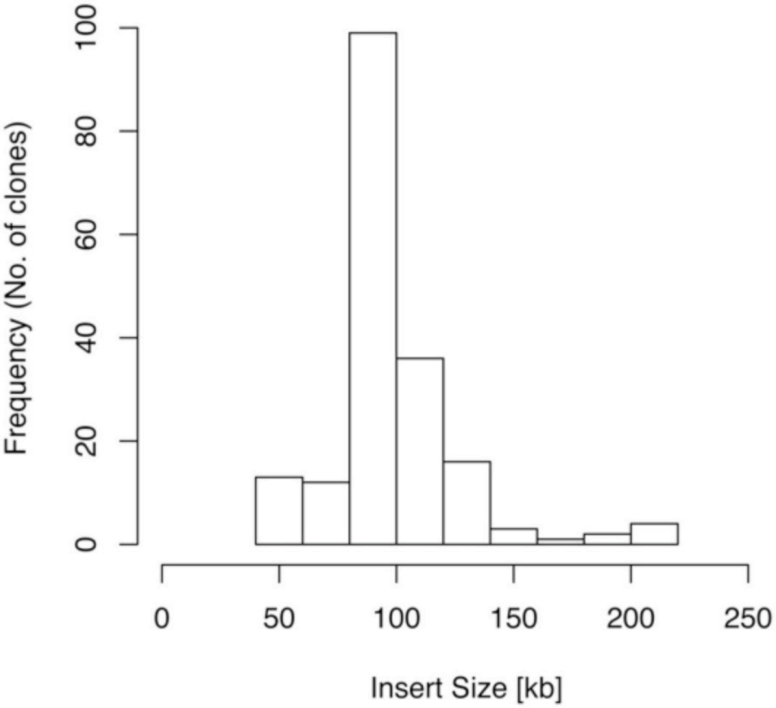
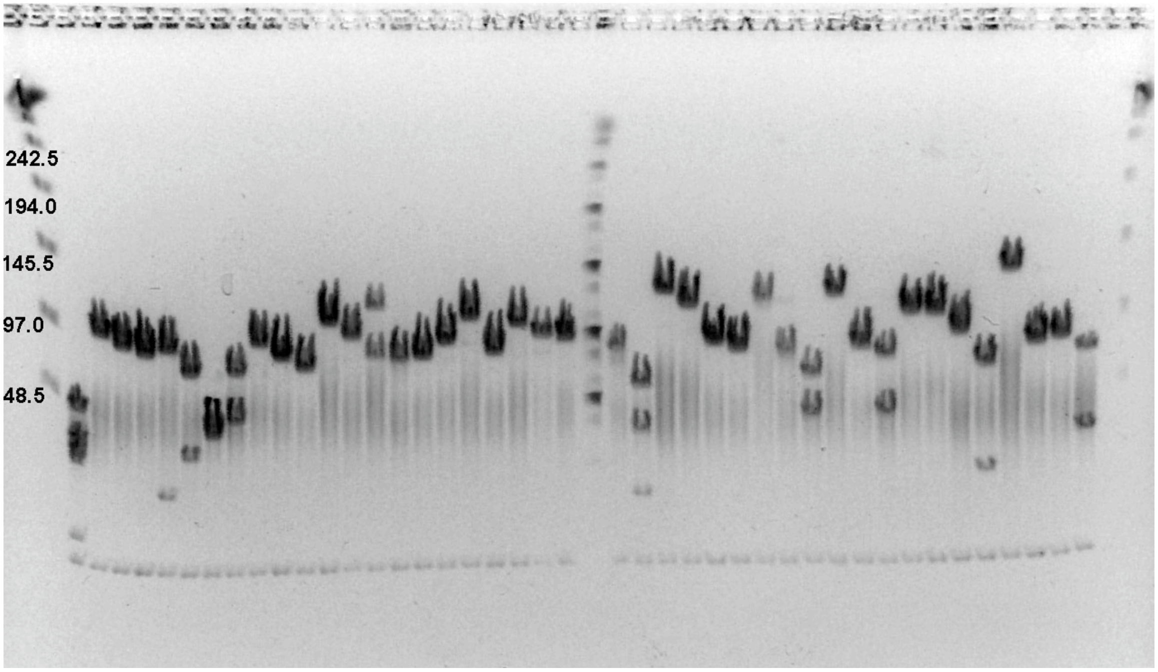


Figure 2.2. Analysis of BAC clones by PFGE. Randomly picked recombinant BAC clones from our *Bombus terrestris* library were digested with NotI to release the cloned genomic insert. Sizes were separated on a 1% agarose CHEF gel (0.5X TBE) and stained with Ethidium bromide. This gel shows the results for 42 BAC clones; the marker loaded in either end-well is Lambda Ladder (NEB). The marker in lane 23 is Midrange II (NEB). Fragment sizes for Lambda Ladders are indicated in Kb on the left.



References

Baer, B and P Schmid-Hempel. 2001. Unexpected consequences of polyandry for parasitism and fitness in the bumblebee, *Bombus terrestris*. *Evolution* **55**:1639-1643.

Beye, M, M Hasselmann, MK Fondrk, RE Page and SW Omholt. 2003. The gene csd is the primary signal for sexual development in the honey bee and encodes an SR-type protein. *Cell* **114**:419-429.

Chen, MS, G Presting, WB Barbazuk, JL Goicoechea, B Blackmon, FC Fang, H Kim, D Frisch, YS Yu, SH Sun, S Higingbottom, J Phimphilai, D Phimphilai, S Thurmond, B Gaudette, P Li, JD Liu, J Hatfield, D Main, K Farrar, C Henderson, L Barnett, R Costa, B Williams, S Walser, M Atkins, C Hall, MA Budiman, JP Tomkins, MZ Luo, I Bancroft, J Salse, F Regad, T Mohapatra, NK Singh, AK Tyagi, C Soderlund, RA Dean and RA Wing. 2002. An integrated physical and genetic map of the rice genome. *Plant Cell* **14**:537-545.

Chew, JS, C Oliveira, JM Wright and MJ Dobson. 2002. Molecular and cytogenetic analysis of the telomeric (TTAGGG)n repetitive sequences in the Nile tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). *Chromosoma* **111**:45-52.

Cook JM and RH Crozier. 1995. Sex determination and population biology in the Hymenoptera. *Trends Ecol Evol* **10**:281-286.

Gadau, J, CU Gerloff, N Kruger, H Chan, P Schmid-Hempel, A Wille and RE Page. 2001. A linkage analysis of sex determination in *Bombus terrestris* (L.) (Hymenoptera : Apidae). *Heredity* **87**:234-242.

Goulson, D. 2003. Bumblebees - their Behaviour and Ecology. Oxford University Press, New York 235p.

Graveley, BR, A Kaur, D Gunning, SL Zipursky, L Rowen and JC Clemens. 2004. The organization and evolution of the Dipteran and Hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA-Publ RNA Soc* **10**:1499-1506.

Hasselmann, M, MK Fondrk, RE Page and M Beye. 2001. Fine scale mapping in the sex locus region of the honey bee (*Apis mellifera). Insect Mol Biol* **10**:605-608.

Heimpel, GE and JG de Boer. 2008. Sex determination in the Hymenoptera. *Annu Rev Entomol* **53**:209-230.

The Honey Bee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honey bee *Apis mellifera*. *Nature* **443**:931-949.

Hoffman, EA and MAD Goodisman. 2007. Gene expression and the evolution of phenotypic diversity in social wasps. *BMC Biol* **5**:23.

Hunt, GJ and RE Page. 1994. Linkage Analysis of Sex Determination in the Honey Bee (*Apis mellifera*). *Mol Gen Genet* **244**:512-518.

Kawakita, A, T Sota, JS Ascher, M Ito, H Tanaka and M Kato. 2003. Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). *Mol Biol Evol* **20**:87-92.

Kawakita, A, T Sota, M Ito, JS Ascher, H Tanaka, M Kato and DW Roubik. 2004. Phylogeny, historical biogeography, and character evolution in bumble bees (*Bombus* : Apidae) based on simultaneous analysis of three nuclear gene sequences. *Mol. Phylogenet Evol* **31**:799-804.

Lazo, GR, N Lui, YQ Gu, XY Kong, D Coleman-Derr and OD Anderson. 2005. Hybsweeper: a resource for detecting high-density plate gridding coordinates. *Biotechniques* **39**:320, 322, 324.

Mahtani, MM and HF Willard. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res* **8**:100-110.

Mardulyn, P. and SA Cameron. 1999. The major opsin in bees (Insecta : Hymenoptera): A promising nuclear gene for higher level phylogenetics. *Mol Phylogenet Evol* **12**:168-176.

Marra, MA, TA Kucaba, NL Dietrich, ED Green, B Brownstein, RK Wilson, KM McDonald, LW Hillier, JD McPherson and RH Waterston. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**:1072-1084.

Pereboom, JJM, WC Jordan, S Sumner, RL Hammond and AFG Bourke. 2005. Differential gene expression in queen-worker caste determination in bumble-bees. *Proc R Soc B-Biol Sci* **272**:1145-1152.

Peterson, D, J Tomkins, D Frisch, R Wing and A Paterson. 2000. Construction of bacterial artificial chromosome (BAC) libraries: an illustrated guide. Published with permission from CAB International. *J Agric Genomics* **5** http://wheat.pw.usda.gov/jag/

Ruiz-Gonzalez, MX and MJF Brown. 2006. Males vs workers: testing the assumptions of the haploid susceptibility hypothesis in bumblebees. *Behav Ecol Sociobiol* **60:**501-509.

Sadd, BM and P Schmid-Hempel. 2006. Insect immunity shows specificity in protection upon secondary pathogen exposure. *CB* **16**:1206-1210.

Sambrook, J, E Fritsch, and T Maniatis. 1989. Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 1659p.

Schmid-Hempel, P. 2001. On the evolutionary ecology of host-parasite interactions: addressing the question with regard to bumblebees and their parasites. *Naturwissenschaften* **88**:147-158.

Soderlund, C, S Humphray, A Dunham and L French. 2000. Contigs built with fingerprints, markers, and FPCV4.7. *Genome Res* **10**:1772-1787.

Tomkins, J, M Luo, G Fang, D Main, J Goicoechea, M Atkins, D Frisch, RE Page, E Guzman-Novoa, Y Yu, GJ Hunt and R Wing. 2002. New genomic resources for the honey bee (*Apis mellifera* L.): development of a deep-coverage BAC library and a preliminary STC database. *Genet Mol Res* **1**:306-316

Wang, J, S Jemielity, P Uva, Y Wurm, J Graff and L Keller. 2007. An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. *Genome Biol* **8**(1):R9.

Wilfert, L, J Gadau, B Baer and P Schmid-Hempel. 2007a. Natural variation in the genetic architecture of a host-parasite interaction in the bumblebee *Bombus terrestris*. *Mol Ecol* **16**:1327-1339.

Wilfert, L, J Gadau and P Schmid-Hempel. 2006. A core linkage map of the bumblebee *Bombus terrestris. Genome* **49**:1215-1226.

Wilfert, L, J Gadau and P Schmid-Hempel. 2007b. The genetic architecture of immune defense and reproduction in male *Bombus terrestris* bumblebees. *Evolution* **61**:804-815.

Yuan, G, C Qi, P Junsong, L Zheng, H Huanle, W Aizhong, S Rentao and C Run. 2008. Construction of a BAC library from cucumber (*Cucumis sativus* L.) and identification of linkage group specific clones. *Prog Nat Sci* **18**:143-147.

CHAPTER THREE

DEVELOPMENT OF BAC LIBRARY RESOURCES FOR PARASITIC HYMENOPTERA (*Nasonia vitripennis* AND *Nasonia giraulti*. PTEROMALIDAE).

Prepared for submission as part of a collection of companion papers to the publication of the *Nasonia vitripennis* Genome Sequence. *Insect Molecular Biology*. In preparation with coauthors Christopher Saski[1], Barbara Blackmon[1], Jeanne Romero-Severson[2], and John H. Werren[3,4].

Authors and Affiliations

Munoz-Torres, M[1], C Saski[1], B Blackmon[1], J Romero-Severson[2], JH Werren[3,4].

[1] Clemson University Genomics Institute. Clemson, SC. 29634. U.S.A.

[2] Department of Biological Sciences. University of Notre Dame. South Bend, IN. 46556. U.S.A.

[3] Department of Biological Sciences. University of Rochester. Rochester, NY. 14627. U.S.A.

[4] Contact information for BAC library resources and data respository.

Abstract

The species of the genus *Nasonia* possess qualities that make them excellent candidates for genetic and genomic studies. To increase the wealth of genomic resources available for these organisms we constructed publicly available BAC libraries for *Nasonia vitripennis* and *Nasonia giraulti*. Each library contains 36,864 clones, average insert sizes estimated at 113.1 Kb for *N. vitripennis* and 97.7Mb for *N. giraulti* representing approximately 12 and 11 genome equivalents respectively, and empty-vector contents of approximately 2%. Additionally, we describe preliminary results on two research projects undertaken with the use of the *N. vitripennis* library. The first one, reports on steps taken towards positional cloning of a gene believed to affect wing size differences among three of the sibling species. The second is a report on a preliminary survey of the *N. vitripennis* genome, obtaining over 1400 BAC-end sequences.

Introduction

Improving our wealth of knowledge about the structure and function of the genome of our organism of interest requires a set of crucial and necessary steps, the first of which is the development of tools that allow researchers to survey the genome. Genomic resources are utilized in a wide array of genetic studies ranging from projects such as the identification of individual genes responsible for an organism's response to changing environmental conditions, or the characterization of genes for improvement of certain traits of agronomic importance in staple crops, all the way to functional studies on a genome-wide scale. Examples of the broad range of answers that may be obtained with the implementation of genomic resources include i) positional cloning of a gene partially responsible for structural changes in free-living adults of pacific oyster (*Crassostrea gigas*) upon fixation to a surface (Unpublished data, Andrew Mount, Department of Biological Sciences, Clemson University), ii) isolation of the gene underlying the sex determination locus in honey bee (Beye et al., 2003, Tomkins et al., 2002), iii) identification of candidate genes for drought-stress tolerance (Xu-Sheng et al., 2006) and iv) disease resistance (Ronald, 1997) in rice (*Oryza sativa* L), v) identification of genes encoding seed storage proteins in quinoa (*Chenopodium quinoa* Willd) (Stevens et al., 2006) and vi) genome mapping in cassava (*Manihot esculenta* Crantz) as a way to identify candidate genes to overcome micronutrient malnutrition challenges worldwide (Fregene et al., 2001). Such diverse research projects have in common the use of a Bacterial Artificial Chromosome (BAC) Library as a way of interrogating the genome in search for practical answers to each problem.

Initially regarded as a system that could facilitate the construction of DNA libraries of complex genomes with fuller representation and subsequent rapid analysis of complex genomic structure (Shizuya et al., 1992), BAC libraries are nowadays a common entry point into projects involving functional and structural genomics, gene identification and construction of physical maps, to name a few. Here we describe the construction and characterization of BAC libraries for two species of parasitic wasps of the genus *Nasonia*, namely *N. vitripennis* (Jewel wasp) and *N. giraulti*.

Parasitic wasps have been subjects of genetic, ecological, evolutionary and developmental research for almost 50 years; they belong to a large and extremely important group of Hymenoptera with more beneficial insects to humans than any other group. *Nasonia* is an excellent candidate for studies in genetics and genomics due to ease of insect rearing, small genome size, haplodiploidy, a wealth of visible and molecular markers available for mapping and healthy inbred isogenic lines. There are four closely related and interfertile species in the genus, the cosmopolitan *N. vitripennis*, and the North American *N. longicornis*, *N. giraulti* (Darling and Werren, 1990) and *N. oneida* (Unpublished data, R. Raychoudhury. Department of Biological Sciences, University of Rochester.), NV, NL, NG and NO respectively and hereafter. The estimated genome size of NV is approximately 335 Mb (Beukeboom and Desplan, 2003 and Rasch et al., 1975) (about two times greater than that of *Drosophila melanogaster*); however, the recombination rate in *Nasonia* is approximately four times greater than in *D. melanogaster*, resulting in an average recombination rate per kilobase approximately two times greater (around 410 Kb/cM (Gadau et al., 1999)). These features grant *Nasonia* the potential to become a great model organism for several areas of research, and the development of genomic tools for this species group will offer an increased amount of resources to exploit the possibilities.

The large diversity of extant insect forms and life styles has made them subjects for phylogenetic, evolutionary and ecological studies. Research on insects has important societal, as well as scientific benefits, as many insects pose serious threats to public health and commercial crops. Despite all of this, insect genomics has remained an uneven territory dominated by studies in Drosophilids and mosquitoes. EST libraries and genetic linkage maps have been constructed for several insects, but BAC libraries are available for only a few species; furtheremore, other than the well-developed genetic model *D. melanogaster*, genomic resources for the remainder of the insects have long been surprisingly inadequate. The availability of large-insert genomic libraries has propelled advances in gene discovery and genome-wide analysis for a reduced number of insect species. BAC libraries are now available for social hymenopterans, honey bee (*Apis mellifera* L.) (Tomkins et al., 2002) and bumblebee (*Bombus terrestris* L.) (Wilfert et al., 2008), the mosquitoes *Aedes aegypti* (Jiménez et al., 2004) and *Anopheles gambiae* (Hong et al., 2003), the red flour beetle *Tribolium castaneum* (Brown et al., 2002) and the silk moth *Bombyx mori* (Mita et al., 2002). Except for bumblebee, genomes of all aforementioned species have now been sequenced, and BAC libraries were used to build preliminary physical maps or to provide sequence scaffolds to facilitate the assembly process in these projects. In a similar manner, both large-insert genomic libraries reported here will offer a starting point for the design of genomic projects for the genus *Nasonia*, and they constitute the first of their kind in the parasitic wasps.

In support of this, federal resources were allocated to develop various genomic tools to exploit this organism. This report describes one of the key resources now publicly available for genomic studies, two NSF funded deep-coverage BAC libraries. In addition, the NIH selected NV for genome sequencing to produce a draft assembly (http://www.genome.gov/10002154). The Human Genome Sequencing Center (HGSC) at Baylor College of Medicine (BCM) has sequenced the genome of *Nasonia vitripennis* at six-fold sequence coverage, and a rough draft of the NG genome has also been produced (Personal communication. Stephen Richards, HGSC, BCM). At the time of this publication, version 1.0 of the assembly, Nvit_1.0, was available for download from the Baylor website. These BAC library resources provide a useful complement to the draft of sequence for a variety of genomic and genetic applications. The development of genomic resources for a parasitoid is also likely to yield many benefits for human health and our understanding of important biological processes.

Results

BAC Library Construction and Characterization

NASONIA VITRIPENNIS. The NV BAC library consists of 36,864 clones. PFGE analysis of 384 randomly sampled clones allowed estimation of insert size at an average of 113.1 ± 39Kb (Figure 3.1) with a range of 8 – 300Kb. Less than 2% of the clones contained no inserts. Based on an estimated haploid genome size of 335Mb (Beukeboom and Desplan, 2003 and Rasch et al., 1975), this library represents 12.4 genome equivalents and allows any one particular NV sequence to be recovered with a probability greater than 99% from at least one clone.

NASONIA GIRAULTI. The NG BAC library counts with 36,864 clones. One hundred and ninety two clones were randomly selected and analyzed using PFGE to estimate library insert size and other features. Based on an estimated genome size of 330Mb (Unpublished data. John H. Werren, University of Rochester) our library covers the NG genome 10.9 times with an estimated average insert size of 97.7 ± 25.9Kb (Figure 3.2) and a range of 9 – 135Kb. Such coverage allows any particular species specific sequence to be recovered with a probability greater than 99% from at least one clone. Approximately 3.1% of the clones contained no inserts.

BAC Library Screening and Fingerprinting Analysis

Its natural history and genetic features, described above, grant *Nasonia* the potential to become a model organism for a number of studies ranging from pharmaceutical uses to increasing our knowledge in fundamental evolutionary biology. In order to test library coverage and isolate genomic regions putatively associated with genes of interest, we screened the BAC libraries with probes derived from candidate genes involved in evolution of wing cell size (*ws1*), insulin regulation pathway (*target of rapamycin* (*tor*) and *S6 kinase* (*S6k*)), tumor suppressing (*phophatase and tensin homolog* (*Pten*)) and embryonic development (*zernkult* (*zen*), *hairy* (*hry*), *caudal* (*cad*) and

Antennapedia (*Antp*)). Table 3.1 contains a list of all primers used to generate these probes.

NG males have large wings and fly, NL males have intermediate wings, and NV males have small vestigial wings and are capable of limited flight. A QTL analysis revealed four to five genetic regions responsible for most of these differences (Weston, 1999). *ws1* is a QTL known to cause a 50% increase in wing cell size in *Nasonia* males and has evolved in NG. To further characterize our two BAC libraries we engaged in selecting a set of candidate BAC clones, which would presumably contain the gene or the tightly linked genes responsible for this trait. We developed a probe from an AFLP fragment, *AF-1*, located within QTL interval of *ws1* in NG. Screening of NV BAC library filters yielded 19 positive hits, which were fingerprinted using HindIII; after manual analysis to filter out poorly resolved fingerprints, 17 clones were assembled into one contig using FPC. BAC end sequences of external most clones in the contig were used to obtain PCR products, which were also labeled to screen the BAC filters again. The probes used were labeled as *67D12.gg*, *01N08.bb*, *02K01.sp* and *19G07.sp* (Table 3.1, Figure 3.3).

Thirty clones were brought together into a contig (Contig Nv1) which included 18 hits from our initial hybridization experiment After two rounds of 'end-probes' hybridizations and analysis, we narrowed down the area containing the homolog (or homologs) responsible for *ws1* in NV to a genomic region spanning approximately 150 - 200Kb between the *AF-1*marker and the outer-most end of clone 19G07, with respect to Contig Nv1 (Figure 3.3). Using the 'Minimum Tile' function in FPC we chose three BAC clones, which best represented this region. Using the original PCR primers we corroborated the presence of *AF-1* and other corresponding ‘end-probes’. Clones 01N08, 02K01 and 52F17 were selected for sequencing.

To select candidate BACs for *ws1* from NG we first identified the NG region corresponding to the genomic region around *ws1* in NV. To do this we screened the library with the *AF-1*, *01N08.bb* and *02K01.sp* probes, and obtained a total of 10, 19 and 12 positive hits, respectively. After FPC analysis, *AF-1* and *01N08.bb* hits formed a single contig (Contig Ng1) with 23 clones, and 7 of the 12 *02K01.sp* hits went into a separate contig (Contig Ng2). One possible explanation for this is that a smaller average insert size in the NG library might have impeded our efforts of bridging from BAC to BAC using the ‘*ws1’* probes previously generated. Despite obtaining positive hits with all probes, it is possible that we are still faced with a gap and some parts of this region could be missing. Using the ‘Minimum Tile’ function in FPC we chose a set of three candidate BACs for sequencing. Clones 47D01 and 89M15 best spanned Contig Ng1, while clone 61B24 spanned almost all of Contig Ng2, made up solely of *02K01.sp* hits spanning approximately 210Kb of genomic DNA. All sequences were deposited to GenBank and identification numbers were assigned as shown in Table 3.1. Further annotation of all putative genes in the selected regions will provide data necessary to fully characterize *ws1* in both species.

In similar experiments, seven other gene probes were used to assess library coverage of the NV BAC library. Our goals were also to isolate additional regions of particular interest and to choose another set of candidate BACs for sequencing as a way of highlighting the relevance of the availability of new genomic resources for the Hymenoptera research community. Excluding *Pten* and *Antp*, probes consisted of PCR products obtained with primers designed directly from *Drosophila melanogaster* sequences (see Table 3.1). The *Pten* and *Antp* probes were designed with degenerate primer approaches described in Baudry et al., 2006 and Munoz-Torres et al., *In preparation, Chapter IV*. As done with previously described probes, we used approximately 30uCi of radioactive phosphorus to label approximately 60ng of probe DNA in each hybridization experiment. We obtained a total of 12 positive hits with the *tor* probe, 5 positive hits with *S6K*, 27 with *Pten*, 8 hits with *zen*, 4 for *hry*, 12 for *cad* and 11 positive hits with probe DNA from the *Antp* gene. All seven genes are known to be present in single copy in the *D. melanogaster* genome according to FlyBase (Wilson et al., 2008). These, and results obtained using the *ws1* probe with both *Nasonia* species, are good indicators of the redundancy of the libraries. After insert size verification of all positive hits, clones with the largest inserts from the NV library were selected as candidate BACs for full length sequencing to make these resources available for further gene characterization in this species. Clone sequences were deposited in GenBank and identification tags are also shown in Table 3.1.

BAC end sequencing

We chose a random sample of clones for end-sequencing and annotation to perform a preliminary survey of the NV genome. The number of high-quality sequences was 1,217, with an average high-quality base count of 486bp. The highest quality match for each sequence tag connector (STC) with a probability cutoff value (E value) of at least 10e-6 was used to assign putative identities to the STCs. The results from our FASTX searches and STC analysis can be obtained from our website at: http://www.genome.clemson.edu/downloadData/nasonia/NV__Bb.fasta.

A BLAST search against the non-redundant collection of protein sequences in GenBank resulted in 178 (14.63%) of the sequences showing similarity to genes of known function. Significant search results were then sorted into seven different functional categories (Table 3.2). The largest group of hits belongs to the hypothetical or unclassified category representing 28% of the dataset. The next largest group (27%) shared sequences similar to insertion elements and encoded maturases, transposases, and integrases. Thirdly, 20% of the STCs are involved in regulatory roles, followed by metabolism (15%), cell communication/division (7%), structural roles (2%), and ribosomal proteins (0.5%). Of the highly significant STCs showing similarity to various proteins, 29 % were best matches to *Apis mellifera* proteins, 12 % shared matches to *Drosophila buzzati*, *D. melanogaster*, or *D. pseudoobscura*, and 6% were best to *Anopheles gambiae*. The rest of the hits were to a wide variety of eukaryotic organisms.

Discussion

We here describe the construction and characterization of BAC libraries for two species of parasitic wasps, *N. vitripennis* and *N. giraulti.* These deep coverage libraries are now available for research projects involved in individual gene isolation as well as functional genome-wide scale experiments. We screened our libraries with probes generated from genes involved in a diverse arrange of biological processes; we retrieved a number of candidate BACs putatively containing these genes of interest demonstrating the high quality and usefulness of our libraries. Candidate BACs from both species were also chosen for full-length sequencing to provide resources for further characterization of such genes of interest. Choosing candidate BACs for sequencing also provided data for assembly verification performed by the *Nasonia* Genome Sequencing Consortium (The *Nasonia* Genome Sequencing Consortium, Unpublished data.).

Identification of candidate genes responsible for variations in a certain trait of interest requires access to the physical genome, and not only to recombination rates. Combining both resources allows for the rapid identification of candidate genes. Our libraries have offered valuable genomic resources for the identification of a candidate gene (or group of tightly linked genes) responsible for a large wing size difference between males of two closely related species of the parasitoid wasp *Nasonia*. In addition to demonstrating the good quality of our libraries, this experiment also presents *Nasonia* as a good model system for genetic analyses of morphological differences between natural populations or closely related species. Traditionally these studies are posed with difficulties due to either small or subtle differences, or the impossibility to perform crosses, even between closely related species. However, in the case of *Nasonia*, the four species described in this genus are interfertile when cured of their cytoplasmic bacterial infections (*Wolbachia*), and they are able to form viable and fertile hybrids (Breeuwer and Werren 1995 and 1990, R. Raychoudhury, Unpublished data).

We carried a preliminary survey of the NV genome, choosing random BAC clones for end sequencing. The results of these experiments represent the first genome-wide survey of NV. That only a small portion of annotated sequences showed similarity with genes of known function (15% of all sequences) may be due to the random process used to choose BACs for end-sequencing. Nonetheless, these results constitute additional demonstration of the high quality of the library and its usefulness. The second largest group of annotated end-sequences contained several proteins with high similarity to sequences of viral origin including insertion elements and transposases; these sequences were also commonly found in the honey bee genome, the only hymenopteran genome sequence finished to date (The Honey Bee Genome Sequencing Consortium, 2006). As expected, the vast majority of annotated sequences with highly significant similarity to known genes were associated with sequences from *A. mellifera*, *A. gambiae* and species of Drosophila, adding confidence to the measure of quality of the genomic resources here introduced.

Tractability and all other features described for the species of the genus *Nasonia* have made them a primary model for parasitoid genetics (Whiting, 1967). More is known about the biology of *Nasonia* than any other parasitic hymenopteran, and an incredibly active research community will ensure that all resources generated for this group of species, such as the two BAC libraries presented in this report, are exhaustively utilized for advanced research.

The clones and filters from the *Nasonia vitripennis* and the *Nasonia giraulti* BAC libraries are publicly available and may be ordered from the Clemson University

Genomics Institute (http://www.genome.clemson.edu/). The use of these BAC libraries should make reference to this resource.

Experimental Procedures

BAC Library Construction

DNA for library construction was extracted from highly inbred lines of NV and NG. For both species, high-molecular-weight DNA from yellow pupae was prepared following a procedure adapted for honey bees (Tomkins et al., 2002). DNA was partially digested with HindIII for 20 minutes at 37°C. Size selection was performed on the fragments via two consecutive rounds of pulse field electrophoresis (PFGE). Fragments were then ligated into the vector pIndigoBAC 536 (Peterson et al., 2000, Luo and Wing, 2003) for 16 hours at 16°C. Vectors were transformed into *E. coli* ElectroMAX DH10B cells (Invitrogen, Carlsbad, CA, USA) using electroporation, and allowed to replicate at 325rpm for 1h at 37°C. Recombinant colonies were picked using a Genetix Q-bot (Genetics, Boston, MA, USA) and stored individually in 384-well plates at -80°C.

BAC Library Characterization

DNA from BAC clones was prepared from 900 uL cultures of Terrific Broth (GIBCO)-Chloramphenicol (12.5 ug/uL) in 96-well format, inoculated with 1.5uL of BAC culture. After 18h, cultures were treated with a modified alkaline lysis method. To determine insert sizes, their distribution, and percent of clones without an insert, approximately 200 ng of BAC DNA from randomly selected clones were digested using

7 units of NotI for 16h at 37°C. DNA digestions were analyzed by PFGE in 1% agarose gels for 15h at 14°C (6 v/cm, switch time of 5-15s) and stained with Ethidium Bromide for 20m.

BAC Library Screening

High-density colony filters for hybridization-based screening of the NV and NG libraries were prepared using a Q-bot (Genetics, Boston, MA, USA). Clones were gridded in double spots using a 4 x 4 array with 6 fields on 11.5 x 22.5 cm Hybond N+ filters (GE Healthcare, Piscataway, NJ, USA). This gridding pattern allows 18,432 clones to be represented per filter. Colony filters were grown and processed using standard techniques (Sambrook et al., 1989). Hybridization probes were designed by cutting PCR fragments from ethidium bromide-stained 1% agarose gels and DNA was extracted using a QIAEX II gel extraction kit (QIAGEN, Valencia, CA, USA). The NV library was screened with eight PCR products obtained from genes involved in evolution of wing cell size, insulin regulation, or embryonic development (i.e.: segment formation and Hox genes). The NG library was screened with the probe developed for the wing cell size gene and subsequent BAC-end probes designed for this experiment, as described in the results section. Radio labeling of probe DNA and hybridization of colony filters were performed using standard techniques (Sambrook et al., 1989) with the following modifications: hybridizations were performed for at least 16h at 65ºC and filters were washed twice at 65ºC (30 min per wash) in a 1X SSC/0.1% SDS solution first, and a 0.5X SSC/0.1% SDS solution the second time. Preparation of single stranded radio labeled DNA probes from

PCR products was done according to manufacturer's instructions using the Random Primed DNA Labeling Kit DECAprime II (Applied Biosystems, Foster City, CA, USA). Hybridized BAC filters were imaged in a Storm Scanner (GE Healthcare, Piscataway, NJ, USA) and positive hits were scored with HybSweeper (Lazo et al., 2005).

Fingerprinting Analysis

BAC-DNA fingerprinting was performed using techniques established by Chen et al., 2002 and Marra et al., 1997. Briefly, BAC DNA was prepared as described above and samples for fingerprinting were digested with the restriction endonuclease Hind III, electrophoresed on 1% agarose gels for 15h at 60V, and stained with SybrGold (Invitrogen, Carlsbad, CA, USA) for 1h. Gels were imaged in a Storm Scanner (GE Healthcare, Piscataway, NJ, USA) and fingerprinting data were scored using Image3 (v.3.10, www.sanger.ac.uk/software/Image/). All bands were manually checked; bands below 1Kb were ignored due to bad image resolution; these accounted for <1% of the total length of BAC inserts. Contig building was done using FingerPrinted Contigs (FPC v.8; Souderland et al., 2000) at a high stringency, with a Tolerance value of 7 and a Minimum Cutoff value of 1e-9. Poorly resolved fingerprints, that is, clones with band patterns of 4 or less bands, were automatically excluded by Image3. FPC automatically excluded those clones with band patterns not significantly similar to any other clone in the data set from contig analysis.

BAC End Sequencing

Preparation of high quality BAC DNA for end sequencing was done in a 96-well format using the standard alkaline lysis miniprep techniques described for BAC-DNA fingerprinting with some modifications. The difference lies in the use of 1.2mL of 2XYT (GIBCO)-Chloramphenicol (12.5 ug/uL) cultures. Sequencing was performed using a dye-terminator cycle sequencing kit (Perkin Elmer Applied Biosystems, USA). Reactions were performed in one direction with T7 on 1344 reads in total reaction volumes of 25uL. Reactions were separated on an ABI 3730xl DNA analyzer (Applied Biosystems, Foster City, CA, USA) for 90 min with an injection time of 15 seconds. Cycle sequencing reactions were done on PTC-200 Thermalcyclers (MJ Research) in 96-well format with the following parameters: 95ºC for 4min followed by 75 cycles of 95ºC for 15sec, 51ºC for 10sec and 60º C for 4min. Ethanol precipitation was applied to remove excessive terminators from sequencing reactions and purified reactions were resuspended in 8uL of HiDi-Formamide (Applied Biosystems, Foster City, CA, USA). Base-calling was performed using Phred (Ewing & Green, 1998; Ewing et al., 1998) and vector sequences were removed using CROSS-MATCH (Green, 1999). High quality sequences are those defined as having at least 100 high quality bases (greater than PHRED 20) other than *E. coli* or vector. These sequences were used as queries in searches against GenBankNR and SWISS-PROT databases using FASTX3.4 algorithms. Software was locally run on either a Macintosh G5 or SunBlade workstations using Solaris (v.9).

Acknowledgements

The construction and characterization of the resources presented in this report were partially funded by a grant from the U.S. National Science Foundation (IOS. 0208278) to J. Romero-Severson. MM-T wishes to thank Dr. Amy Lawton-Rauh for helpful reviews on this manuscript. MM-T, CS, BB and JT wish to thank Mrs. Jeanice Troutman and Mr. Michael Atkins for technical assistance.

Table 3.1. Probes designed for hybridization of the *N. vitripennis and N. giraulti* BAC libraries. Positive clones were fingerprinted using HindII and assembled into contigs. Clones with largest inserts were selected for complete sequencing.

Gene[*]	Primers (5'-3')	Source Organism	BAC Address[+]	BAC GenBank ID
antp	Degenerate primers	Degenerate primers	NV__Bb_49K07	M Munoz-Torres[§]
	M Munoz-Torres[§] Chapter IV			
cad	cad-F CATGAATTCAARACKCGNACKAARGAYAARTA	*Drosophila melanogaster*	NV__Bb_66L02	AC185337
Oleniski et al. 2006	cad-R TGAGTCGACRTTYTGRAACCADATYTTNAC			
hry	C. Desplan[§]	*Drosophila melanogaster*	NV__Bb_67B15	AC185339
NM_001014577				
Pten	Pten-F GATGACCACAATCCACCCC	Degenerate primers	NV__Bb_85B05	AC185134
J. Romero-Severson[§]	Pten-R AANGTRTTYAACCARAARTG			
s6k	s6k-F GCATTTCAGACGGAG	*Drosophila melanogaster*	NV__Bb_29I19	AC185290
J. Romero-Severson[§]	s6k-R GAATGTCCTTTCCGG			
tor	tor-F GGAGCTCGACCAGGT	*Drosophila melanogaster*	NV__Bb_44F11	AC185143
J. Romero-Severson[§]	tor-R TTGACCATAAGCTGC			
ws1	ws1-F GTCGCACCACTTGCTACAGA	*Nasonia Vitripennis*	NV__Bb_01N08	AC185338
J. Werren[§]	ws1-R GAATTGGCCAACCTTCATTG		NV__Bb_02K01	AC185141
	01N08.bb-F GGCGAACTGCATACATGCGC		NV__Bb_52F17	AC185288
	01N08.bb-R TCTGAGCAGCTAATAAAGGCAGAGC		NG__Bb_47D01	AC185330
	02K01.sp-F CTACAGTGGTGAGCAGAGGTC		NG__Bb_61B24	AC185331
	02K01.sp-R TCCTTGCGGTACTCTCTTG		NG__Bb_89M15	AC185140
	19G07.sp-F CAGTGATTTTGCACTCGTTC			
	19G07.sp-R TTACCTTTACGGCTTGTTTG			
	67D12.gg-F AGCTTGATGCAGCGTGACTTC			
	67D12.gg-R AAATGTGCGACGAGCTGAGC			
zen	C. Desplan[§]	*Drosophila melanogaster*	NV__Bb_77L07	AC185133
NM_057445				

§ Unpublished data

* Includes abbreviated name, GenBankID and/or reference when available. See text for complete gene names.

+ NV__Bb for *Nasonia vitripennis* BAC library, NG__Bb for *N. giraulti*.

Table 3.2. BLAST results of 178 BAC end sequences from randomly chosen clones of the *Nasonia vitripennis* BAC library.

Functional Category	**Number**	**Percent**
Retroelement	48	26.9
Enzymatic	27	15.1
Regulatory (kinase, phosphatase, transcriptional, transport)	36	20.2
Structural	4	2.2
Ribosomal	1	0.5
Cell defense, communication and division	13	7.3
Hypothetical/unclassified	49	27.5

Figure 3.1. Histogram of insert size distribution of BAC clones (n = 384) of the *Nasonia vitripennis* BAC library.

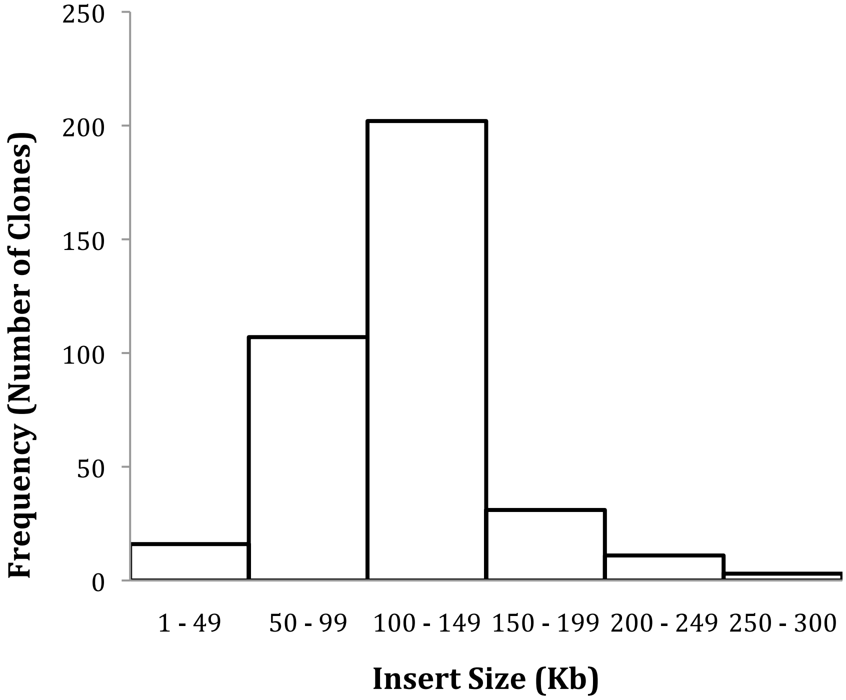


Figure 3.2. Histogram of insert size distribution of BAC clones (n = 192) of the *Nasonia giraulti* BAC library.

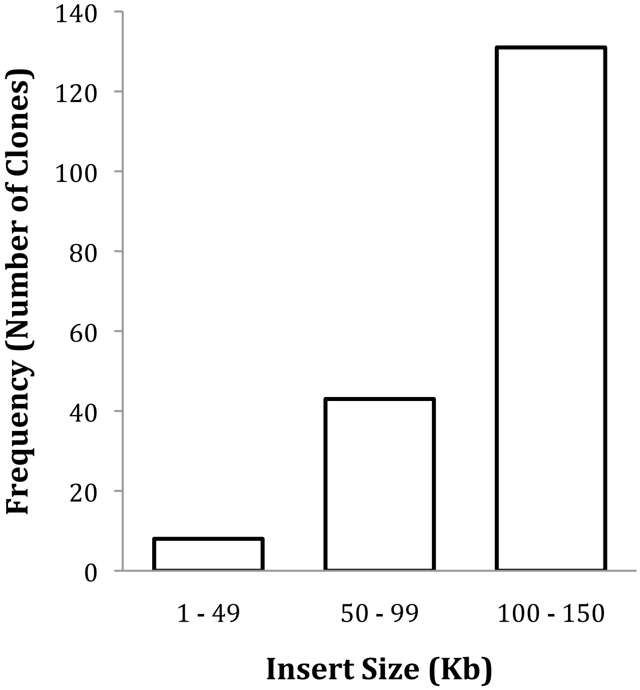
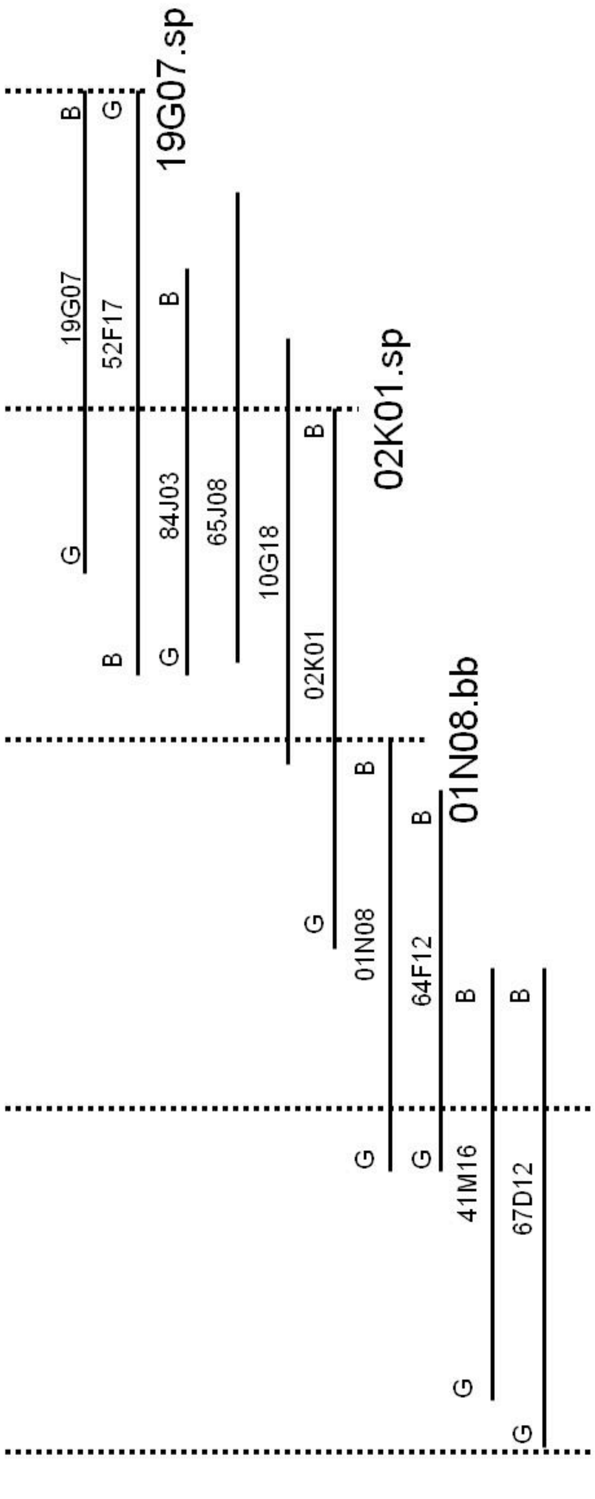


Fig. 3.3. Contig Nv1. Not-to-scale representation of contiged BAC clones after fingerprinting analyses. BACs in this figure were retrieved with subsequent hybridizations starting with the AFLP-derived marker ‘*Af1*’, located near *ws1* and continuing with end-probes *67D12.gg*, *01N08.bb*, *02K01.sp* and *19G07.sp*. ‘G’ marks the T7 end of the BAC. ‘B’ denotes the SP6r end of the BAC. See text for detailed information.


19G07.sp
02K01.sp
01N08.bb
Af1
67D12.gg
19G07
52F17
84J03
65J08
10G18
02K01
01N08
64F12
41M16
67D12
G
B

References

Baudry, E, M Desmadril and JH Werren. 2006. Rapid Adaptive Evolution of the Tumor Suppressor Gene Pten in an Insect Lineage. *J Mol Evol* **62**:738-744.

Beukeboom, L and C Desplan. 2003. *Nasonia*. *CB* **13**(22):R860.

Beye, M, M Hasselmann, MK Fondrk, RE Page and SW Omholt. 2003. The gene csd is the primary signal for sexual development in the honey bee and encodes an SR-type protein. *Cell* **114**:419-429.

Breeuwer, JA and JH Werren. 1990. Micoorganisms associated with chromosome destruction and reproductive destruction and reproductive isolation between two insect species. *Nature* **346**:558-560.

Breeuwer, JA and JH Werren. 1995. Hybrid breakdown between two haplodiploid species: the role of nuclear and cytoplasmic genes. *Evolution* **49**:705-717.

Brown, S, JP Fellersb, TD Shippy, EA Richardson, M Maxwell, JJ Stuart, and RE Denella. 2002. Sequence of the *Tribolium castaneum* Homeotic Complex: The Region Corresponding to the *Drosophila melanogaster* Antennapedia Complex. *Genetics*. **160**:1067-1074.

Chen, M, G Presting, WB Barbazuk, JL Goicoechea, B Blackmon, G Fang, H Kim, D Frisch, Y Yu, S Sun, S Higingbottom, J Phimphilai, D Phimphilai, S Thurmond, B Gaudette, P Li, J Liu, J Hatfield, D Main, K Farrar, C Henderson, L Barnett, R Costa, B Williams, S Walser, M Atkins, C Hall, MA Budiman, JP Tomkins, M Luo, I Bancroft, J Salse, F Regad, T Mohapatra, NK Singh, AK Tyagi, C Soderlund, RA Dean, RA Wing. 2002. An Integrated Physical and Genetic Map of the Rice Genome. *Plant Cell* **14:**537-545.

Darling, DC and JH Werren. 1990. Biosystems of *Nasonia* (Hymenoptera: Pteromalidae): two new species reared from bird's nests in North America. *Ann Ent Soc Am* **83**:352-370.

Ewing, B and P Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**:186-194.

Ewing, B, L Hillier, M Wendl, and P Green. 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Res* **8**:175-185.

Fregene, M, E Okogbenin, C Mba, F Angel, MC Suarez, JP Guitierez, P Chavarriaga, W Roca, M Bonierbale and J Tohme. 2001. Genome mapping in cassava improvement: Challenges, achievements and opportunities. *Euphytica* **120**:159-165.

Gadau, J, RE Page Jr. and JH Werren. 1999. Mapping of Hybrid Incompatibility Loci in *Nasonia*. *Genetics* **153**:1731-1741.

Green, P. 1999. Swat/cross_match/phrap Package. Available at http://www.phrap.org

The Honey Bee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera. Nature* 443:931-949.

Hong, YS, JR Hogan, X Wang, A Sarkar, C Sim, BJ Loftus, C Ren, ER Huff, JLCarlile, K Black, HB Zhang, MJ Gardner and FH Collins. 2003. Construction of a BAC library and generation of BAC end sequences-tagged connectors for genome sequencing of the African malaria mosquito *Anopheles gambiae*. *Mol Gen Genomics* **268**:720-728.

Lazo, GR, N Lui, YQ Gu, XY Kong, D Coleman-Derr and OD Anderson. 2005. Hybsweeper: a resource for detecting high-density plate gridding coordinates. *Biotechniques* **39**:320, 322, 324.

Jiménez, LV, BK Kang, B de Bruyn, DD Lovin and DW Severson. 2004. Characterization of an *Aedes aegypti* bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence Plasmodium susceptibility. *Insect Mol Biol* **13**(1):37-44.

Luo M, RA Wing. 2003. An Improved Method for Plant BAC Library Construction. Plant Functional Genomics. Ed Erich Grotewoldl. Series: Methods in Molecular Biology. **236**:3-19.

Marra, M, K Dewar, P Dunn, JR Ecker, S Fischer, S Kloska, H Lehrach, M Marra, R Martienssen, S Meier-Ewert and T Altmann. 1997. High throughput fingerprint analysis of large-insert clones: Contig construction and selection of clones for DNA-sequencing. *Genome Res* **7**:1072-1084.

Mita, K, M Mitsuoki, O Kazuhiro, K Yoshiko, N Junko, MG Suzuki and T Shimada. 2002. Construction of an EST database for Bombyx mori and its applications. *Current Science* **83**(4):426-431.

Olesnicky, EC, AE Brent, L Tonnes, M Walker, MA Pultz, D Leaf and C Desplan. 2006. A caudal mRNA gradient controls posterior development in the wasp *Nasonia*. *Development* **133**:3973-3982.

Peterson, D, J Tomkins, D Frisch, R Wing and A Paterson. 2000. Construction of bacterial artificial chromosome (BAC) libraries: an illustrated guide. Published with permission from CAB International. *J Agric Genomics* **5** http://wheat.pw.usda.gov/jag/

Rasch, EM, JD Cassidy and RC King. 1975. Estimates of genome size in haploid diploid species of parasitoid wasps. *J Histochem Cytochem* **23**:317.

Ronald, PC. 1997 The molecular basis of disease resistance in rice. *Plant Mol Biol* **35**:179-186.

Sambrook, J, E Fritsch, and T Maniatis. 1989. Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 1659p.

Shizuya, H, B Birren, U J Kim, V Mancino, T Slepak, Y Tachiiri, and M Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci USA* **89**(18):8794-8797.

Souderland, C, S Humphray, A Dunham and L French. 2000. Contigs built with fingerprints, markers and FPC V4.7. *Genome Res* **10**:1772-1787.

Stevens, MR, CE Coleman, SE Parkinson, PJ Maughan, HB Zhang, MR Balzotti, DL Kooyman, K Arumuganathan, A Bonifacio, DJ Fairbanks, EN Jellen, JJ Stevens. 2006. Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. *Theor Appl Genet* **112**:1593-1600.

Tomkins, J, M Luo, G Fang,D Main, J Goicoechea, M Atkins, D Frisch, RE Page, E Guzman-Novoa, Y Yu, GJ Hunt and R Wing. 2002. New genomic resources for the honey bee (*Apis mellifera* L.): development of a deep-coverage BAC library and a preliminary STC database. *Genet Mol Res* **1**:306-316.

Wang, Xu-sheng, J Zhu, L Mansueto, R Bruskiewich. 2005. Identification of candidate genes for drought stress tolerance in rice by the integration of a genetic (QTL) map with the rice genome physical map. *J Zhejiang Univ SCI* **6B**(5):382-388.

Weston, RF, I Qureshi and JH Werren. 1999. Genetics of a wing size difference between two *Nasonia* species. *J Evol Biol* **12**:586-595.

Whiting, AR. 1967. The biology of the parasitic wasp *Mormoniella vitripennis*. [=*Nasonia brevicornis*] (Walker). *Q Rev Biol* **42**:333-406.

Wilfert, L, M Muñoz Torres, C Reber-Funk, R Schmid-Hempel, J Tomkins, J Gadau and P Schmid-Hempel. 2008. Construction and characterization of a BAC-library for a key pollinator, the bumblebee *Bombus terrestris* L. DOI: 10.1007/s00040-008-1034-1.

Wilson, RJ, JL Goodman, VB Strelets, and the FlyBase Consortium. 2008. FlyBase: integration and improvements to query tools. *Nucleic Acids Res* **36**:D588-D593.

CHAPTER FOUR

SELECTION DIFFERENTIAL ACROSS ANTENNAPEDIA AMONG HYMENOPTERAN SPECIES SUGGEST HOMEODOMAIN-SPECIFIC PURIFYING SELECTION

Prepared for submission as part of a collection of "companion papers" to the publication of the *Nasonia vitripennis* Genome Sequence. *Insect Molecular Biology*. In preparation with coauthor Amy Lawton-Rauh[2].

Authors and Affiliations

Muñoz-Torres, M[1] and A Lawton-Rauh[2].

[1] Clemson University Genomics Institute. Clemson, SC. 29634. U.S.A.

[2] Department of Genetics and Biochemistry. Clemson University. Clemson, SC. 29634.

Abstract

Hox genes are a family of homeodomain-encoding transcription factors which activate and repress a plethora of downstream genes as they bind to DNA directly in a sequence-specific fashion. They control segment identity in the developing embryo of all Arthropods and help to establish the anterior-posterior body axis. Insects have maintained the ancestral Arthropod set of 10 orthologs in the cluster. Differences in life history and body plan of two species of social pollinator bees and one species of solitary parasitic wasps may have placed developmental constraints on the evolution of each of these genes and the cluster as a whole. Despite these differences, the majority of *Hox* display a high degree of conservation across hundreds of millions of years. In this study we conducted evolutionary genetic analyses of one Hox gene, *Antennapedia* (*Antp*) in three species of Hymenoptera. We report differential estimates of synonymous and nonsynonymous substitution rates (ω) along two different domains in the *Antp* coding sequence, indicating that selective pressure may not be occurring at a homogeneous rate throughout the gene. We further hypothesized that these rates may also be dramatically different when compared with those of other genes in the cluster, especially those which no longer play a homeotic role in development.

Introduction

Hox genes are homeodomain transcription factors, very well conserved in sequence and expression across the arthropods and other animals (Hughes and Kaufman, 2002a). This group of genes controls segmentation identities of developing animal embryos along the head-to-tail (anterior-posterior) axis. Nearly all bilaterians share a general genetic framework that forms their body structures, although their body forms are very diverse. *Hox* genes play a central role in the make up of this framework. Functional and evolutionary comparative analyses of homeotic genes across diverse organisms have provided important insights concerning the evolution of developmental patterns in the animal body plan facilitating our understanding of animal diversity (e.g.: Angelini and Kaufman, 2005, Deutsch and Mouchel-Vielh, 2003, Hoegg and Meyer, 2005). Thanks to the discovery of clustering of homeotic genes in *Drosophila* (Lewis et al., 1980a, 1980b, Lewis, 1978, Denell, 1994), the DNA binding domain they encode (McGinnis et al., 1984, Gehring, 1987) and the knowledge that homologous homeotic genes are involved in development of both vertebrates and invertebrates (Bachiller et al., 1994, Carroll, 2000), detailed molecular comparisons across large phylogenetic distances have been possible. Although with some variation, *Hox* function is conserved enough to make comparisons based on expression data reasonable, mainly thanks to the available *Hox* phenotypes (Hughes and Kaufman, 2002a). With a few exceptions, *Hox* genes are arranged in such a way that they are expressed in distinct domains along the anterior-posterior axis, which generally correspond spatially and temporally with their location on the chromosome (Kaufman et al., 1990). It is widely accepted that all paralogs of the *Hox* family of genes originated from a single ancestral gene by duplication. These series of duplication events are likely to be the origin of such spatial and temporal arrangement. Despite being the result of tandem duplication events, only a few cases are known where the cluster has been dispersed or the direction of transcription has changed for one or more genes (e.g. Lemmons and McGinnis, 2006), indicating that some type of selection must be acting on the organization of the cluster in order to maintain this arrangement. Further comparison of *Hox* genes across diverse organisms both at the functional and molecular level will provide important insights concerning the evolution of developmental patterns and may also provide a framework for exploring fundamental principles of regulatory gene evolution.

In arthropods, variations on the expression patterns of *Hox* genes and on the regulation of their downstream targets govern the differences along the evolution of segmental specialization (Averof, 2002). These changes include large shifts in their regional domains of expression and the evolution of finer differences in their expression within individual segments. Comparative studies on the evolution of *Hox* genes through the analysis of the molecular evolutionary history shaping the organization and function of individual genes from the cluster are now possible thanks to the availability of genomic data (Wolfe and Li, 2003, Averof 2002). Measurements of the rate of evolutionary changes between orthologs of the same *Hox* gene among species may yield information on the nature of the selective pressure acting upon the cluster. At the protein level, estimates of the ratio of nonsynonymous (*dN*) to synonymous (*dS*) substitution rates ($\omega = dN/dS$) represent an effective way of testing for patterns consistent with natural selection. Lessons learned form other systems show that members of another family of transcription factors, found in animals, fungi and plants, which also contain a highly conserved DNA-biding domain, are evolving under strong purifying selection. Such is the case of a number of MADS-box genes associated with plant dormancy, the DAM genes (Sergio Jimenez et al., submitted). MADS-box genes are also known to have originated and evolved from a single gene through a series of duplication events and, similar to the role of *Hox* genes in animals, a subset of MADS-box genes exists with homeotic capabilities involved in the development of floral organ identity determination (Coen and Meyerowitz, 1991). Changes in both regulatory and non-regulatory regions of MADS-box genes seem to play important roles during evolution of phenotypic identities of floral organs. So too, such changes in *Hox* transcription factors are likely to play important roles during the evolution of the animal body plan.

On the other hand in vertebrates research suggests that positive selection acted on the homeodomain immediately after *Hox* clusters duplications (Lynch et al., 2006). Given that the location of sites under positive selection in the homeodomain suggests that they are involved in protein-protein interactions, their results further suggest that adaptive evolution actively contributed to *Hox*-gene homeodomain functions, at least in some vertebrate orders. More recent studies performed on Atlantic Salmonid fishes, estimated evolutionary rates along this lineage to test whether positive natural selection is acting on the Homeodomain. After two rounds of genome duplications that predate the origin of vertebrates and third fish-specific duplication event, Salmonid fish underwent a fourth round of genome duplication about 25 to 100 Million years ago (Allendorf et al., 1984). Positive selection could not be detected in this lineage (Mungpakdee et al., 2008).

In this study we focused our attention on estimating and comparing rates and rate ratios of synonymous and nonsynonymous substitutions shaping the evolutionary history a Hox gene, *Antennapedia* (*Antp*) in the Insect order Hymenoptera. Our goal was to investigate whether Darwinian selection is currently acting uniquely on the homeodomain of *Antennapedia* in order to expand our knowledge of the evolution of the Hox Cluster in Insects. This research is the first study of its kind for the Arthropods.

In insects, *Antp* is expressed in a restricted domain in the middle of the embryo (thoracic segments T1-T3) playing a role in patterning the thorax (Hughes and Kaufman, 2002a); it is normally required for the development of the thoracic ectoderm and the formation of legs (Struhl, 1982); *Antp* is also involved in the development of the embryonic midgut (Reuter and Scott, 1990), and the peripheral nervous system (Heuer and Kaufman, 1992) and the central nervous system (Hayward et al., 1995). In the honey bee (*Apis mellifera) Antp* shows strong expression in the thorax and later expansion into the abdomen (Walldorf et al., 2000). *Antennapedia* (*Antp*) is one three genes which evolved very distinct functions after a series of gene duplication events in the ancestor of the Arthropods (de Rosa et al., 1999); the other two genes are *Ultrabithorax* and *Abdominal-A.*

It is possible that different parts of a molecule are subject to different selective constraints (Halliburton, 2004). If positive selection causes divergence in a specific feature in which a protein is known to be involved, parts of the protein involved in such feature would be more divergent than the parts that are not. And the same is true for the opposite case; where purifying selection acting in a specific region of the protein would make nearly any change have some effect in protein structure and/or function, almost certainly a deleterious one. In this case too, a different value of *dN/dS* (ω) is expected for the different portions of the protein. For instance in the insulin molecule, which is composed of two chains capable of forming disulfide bonds with each other and a third chain that is spliced out of the proinsulin peptide, replacement substitution rates are lower across former chains than it is for the spliced out chain (Halliburton, 2004).

To test for natural selection differentially acting on different regions of the *Antennapedia* (*Antp*) gene, we set out to investigate whether the portion of (*Antp*) involved in binding DNA in a sequence-specific manner had different *dN/dS* ratios than portions of the protein which are not involved in the process. We here report results on the use of maximum likelihood methods to conduct molecular evolution analyses testing two hypotheses. First, that the homeodomain has a lower *dN/dS* ratio than the rest of the protein by virtue of the selective constraints acting on a motif that binds DNA in a sequence-specific manner. Second, that this ratio has a value smaller than one ($\omega < 1$), consistent with the expected evolutionary footprint of selective pressure purifying deleterious mutations.

To perform these analyses we sequenced the genomic region containing the *Antennapedia* gene in *Nasonia vitripennis* and *Bombus terrestris* implementing a degenerate PCR procedure to generate the species-specific *Antp* probes and chose candidate BACs for shotgun-sequencing. The deduced *Antp* sequences for these two species of Hymenoptera are 352 (*B. terrestris*) and 362 (*N. vitripennis*) amino acids long. In both species the *Antp* gene comprises two exons and one intron. This paper represents the first study of molecular evolution of a developmental gene in the European bumblebee (*B. terrestris*) and a it takes a unique look at the evolutionary history of a *Hox* gene for both *B. terrestris* and the solitary parasitic jewel wasp (*N vitripennis*). The results show that estimates of the ratio of evolutionary rates (ω) are lower along the coding sequence of the *Antp* homeodomain compared with the rest of the protein, which may be the result of differential evolutionary constraints posed by how selected portions of the protein interact with other proteins or with other regions of the genome.

Results and Discussion

Newly deduced Antennapedia *sequences for* Nasonia vitripennis *and* Bombus terrestris*: sequencing, annotation.*

We deduced the genomic sequences for the region containing the *Antennapedia* (*Antp*) gene in the solitary parasitic wasp *Nasonia vitripennis* and the social pollinator *Bombus terrestris* (Nv and Bt, respectively and hereafter). Sequences obtained from BACs from Nv and Bt spanned genomic regions of 180Kb and 130Kb in size, respectively. Using the program FGENESH+ (Softberry Inc, Mount Kisco, NY, USA) we were able to accurately predict the amino acid sequence of the *Antp* gene in both species. FGENESH (Salamov and Solovyev, 2000) is a popular Hidden Markov Model (HMM)-based gene prediction program. FGENESH works through the recognition of different types of exons, promoters and polyA signals and an optimal combination of these features is then found by dynamic programming and a set of gene models is constructed along the queried sequence. FGENESH+ is an FGENESH variant which incorporates information from known homologous proteins of the queried sequence for more accurate gene assembly from predicted exons. To run FGENESH+ predictions with the Nv BAC sequence, the Am amino acid sequence was used as a training peptide. In the case of Bt both the Am protein sequence and Nv predicted peptide were used as reference homologous sequences. Combined with one additional round of homology searches, which were performed manually on NCBI using BLAST, we finalized the manual annotation for both genes. The predicted peptides covered a genomic region of 11,883bp for Bt and 12,086bp for Nv. Using NCBI's Conserved Domain Search engine (Marchler-Bauer et al., 2007) we annotated the homeodomain on the *Antp* deduced protein sequences highlighting the position of each alpha-helix in the helix-turn-helix motif encoded by this conserved domain. Figure 4.3 shows the relative position of the homeodomain and the helices along the length of the nucleotide alignment for the hymenopteran *Antp*.

To optimize the annotation of Nv *Antp* we took advantage of the availability of a draft version (1.0) of the *Nasonia* Genome Assembly (Nvit 1.0. Personal Communication; Stephen Richards, Human Genome Sequencing Center, Baylor College of Medicine and Christine Elsik, Department of Biology, Georgetown University on behalf of The Nasonia Genome Sequencing Consortium [NGSC]) using an Apollo interface (Lewis et al., 2002). This genome annotation viewer also serves as an editing tool for predicted gene models and its implementation allowed us to simultaneously compare our deduced sequence with *Antp* protein sequences from *D. melanogaster*, *A. mellifera* and *Tribolium castaneum*, as well as with EST supporting evidence for this region from *N. vitripennis*, and *Solenopsis invicta*; an additional gene model for the Nv *Antp* gene obtained as the result of a combination of homology searches and *ab initio* predictions performed with NCBI's gene prediction method GNOMON (Nagy et al., 2008) was also used in Apollo. Despite the amount of supporting information available through Apollo, detailed attention had to be paid when deducing the final annotation of *Antp* for Nv. Genomic sequences from NGSC contained an inserted Cysteine (C) residue in position 1,463,219 of SCAFFOLD23 in Nvit 1.0, which shifted the reading frame in the first exon of the predicted *Antp* gene model. After further inspection of 14 high quality reads of genomic sequences from our BAC shotgun-sequencing effort, which spanned this region, we were able to identify the insertion and correct this sequencing error. (Data not shown). Additionally, BAC shotgun sequencing data allowed us to close an overestimated genomic gap of 413 bp in the NGSC sequences (Positions 1468011-1468423, SCAFFOLD23, Nvit 1.0. Personal Communication, NGSC). The actual length of the gap was 281 bp. These results highlight the monumental importance of performing manual annotations to review automated predictions. Regardless of the high fidelity of the sequencing process and the ever increasing efficiency of protein-prediction algorithms, it is the biological knowledge on each predicted peptide what will ultimately confirm any putative gene models as accepted or rejected. Genomic (gDNA), coding (cds) and amino acid sequences for NV and Bt as well as Nv's genomic gap data may be found in *Appendix C* (*Supplementary data C-1.*).

Evolutionary genetic analyses

MULTIPLE SEQUENCE ALIGNMENTS AND NUMBER OF SYNONYMOUS AND NONSYNONYMOUS DIFFERENCES. A multiple sequence alignment prepared with CLUSTAL 2.0.10 (Larkin et al., 2007) was used as input file for DnaSP (Rozas et al., 2003) and is shown in Figure 4.1. The Nv *Antp* sequence contains 30 nucleotides more than those of Am and Bt. After removing alignment gaps, 347 codons were analyzed (1041 sites) with a total of 230 synonymous sites and 811 non-synonymous sites. The number of synonymous and non-synonymous differences (*S-Dif* and *N-Dif*, respectively) per total number of synonymous and non-synonymous sites was estimated using the Nei & Gojobori's substitution model (NG86, Nei and Gojobori, 1986) in DnaSP (Table 4.1). In the coding region of *Antp* in the species of Hymenoptera here examined the majority of non-synonymous differences (*N-Dif*) are concentrated on regions outside the Homeodomain. For example, when doing a pairwise analysis of non-synonymous differences between Am and Bt 22 of the 23 observed differences occur outside the Homeodomain; the same is true for comparisons between Am and Nv and between Bt and Nv. These results indicate that the region outside the homeodomain may have a higher tolerance for non-synonymous substitutions than the portion of the protein involved DNA-binding, perhaps as a result of less evolutionary constraints.

Implementing the Neighbor-Joining method (NJ) in MEGA4 we inferred the evolutionary history among the three *Antp* loci as shown in Figure 4.2. Evolutionary distances were computed using the Kimura 2-parameter method and are presented in number of base substitutions per site. Codon positions included were 1st + 2nd + 3rd +

Noncoding. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). As expected both members of the family Apidae are shown to be more closely related to each other than either is to *Nasonia*.

DnaSP is a program designed for analysis of DNA polymorphisms from nucleotide sequence data and it computes the synonymous and non-synonymous differences between a given pair of sequences as Nei and Gojobori (1986). If more than one nucleotide difference is present between two codons DnaSP considers all possible pathways of substitution with equal probability, deleting those that lead to stop codons. In instances in which codons differ by multiple changes, DnaSP is unable to report θ estimates per site (the total number of mutations (Watterson, 1975)). Our results showed that θ estimates were not reported for our analyses (data not shown), thus we assumed the occurrence of non-synonymous substitution sites with multiple (more than 1) changes per codon and decided to further investigate. The numbers and distributions of these sites were manually recorded by overlapping the nucleotide and amino acid sequence alignments (*Supplementary data C-2.*) and data were plotted across the length of the nucleotide alignment as shown in Figure 4.3. Only one such codon was observed inside the Homeodomain, suggesting also a lower tolerance for non-synonymous nucleotide changes in this region when compared to the rest of the protein.

DnaSP also measures nucleotide diversity values (the average number of nucleotide substitutions per site between two sequences) using the Jukes and Cantor (1969) correction. It is possible that the proportion of differences may be so high that this correction cannot be computed (Rozas et al., 2003). Preliminary analyses with the

‘Synonymous and Nonsynonymous Substitutions’ command in DnaSP indicated that *dS* could not be computed (data not shown). We compared the nucleotide and amino acid sequence alignments in order to check for possible alignment gaps in the nucleotide sequence that did not correspond with gaps at the amino acid level, changing the frame. In fact in a few instances CLUSTAL inserted a gap in the middle of a codon triplet (data not shown), changing the translation of the deduced amino acid. Because we consider that biologically speaking it is more relevant to conserve the amino acid deduced sequences, we manually modified the location of gaps in the nucleotide sequence alignment to agree with the gaps in the amino acid one. This correction improved the outcome of DnaSP analyses and both evolutionary rates (*dN* and *dS*) were successfully reported, as described below.

However, this also sparked our curiosity to further interrogate these genomic sequences in search for biases in codon usage, which could also account for an unusually high number of synonymous substitutions, skewing the data and making it impossible to calculate the Jukes and Cantor correction. It is widely accepted that codon usage is biased amongst organisms and that it reflects the selection on mutated nucleotides over time (Akashi and Eyre-Walker, 1998). The ‘Codon Usage’ command in DnaSP calculated the Effective Number of Codons (ENC, Wright, 1990), Codon Bias Index (CBI, Morton 1993) and Scaled Chi Square (SChi2, Shields et al., 1998) for each species, to measure the extent of nonrandom usage of synonymous codons (Rozas et al., 2003). Maximum bias in terms of ENC results from the use of only one codon per amino acid, resulting in an ENC value of 20. No codon bias then is the result of equal usage of all synonymous codons for each amino acid and an ENC value of 61. CBI measures the deviation from equal use of synonymous codons and a value of 0 represents the uniform use of these whereas the opposite, maximum codon bias, is represented by a value of 1. Similarly, SChi2 values range from 0 (no bias) to 1 (maximum bias).

Results obtained (Table 4.2) suggest a significant bias in codon usage for *Nasonia vitripennis,* where only 31 of the possible 61 codons have been used in *Antp*. A CBI value of 0.766 and an SChi2 value of 0.825 also suggest a bias in codon usage in the Nv sequence. The conservation and evolution of codon usage patterns can be more discriminating in the level of genomic signature difference than nucleotide abundances on the third site within codons (Chantawannakul and Cutler, 2008). Additionally, these signature differences can also be measured and statistically quantified. Thus, although these values are the result of a single gene, they represent a starting point to inquire more about what appears to be an interesting and distinctive feature of the *Nasonia* genome.

ESTIMATING RATES OF EVOLUTION IN THE ANTENNAPEDIA GENE. To estimate the extent of polymorphism and divergence along *Antennapedia* loci from Am, Bt and Nv, the ratio of rates of evolutionary change *ω* (*dN/dS*) was measured with DnaSP. We calculated the rates of synonymous and nonsynonymous substitutions for the complete coding sequence of *Antennapedia* gene among the three species (Am, Bt and Nv) and separately for partial coding sequences covering the homeodomain (hd) and non-homeodomain (nhd) regions. Table 4.3 describes our findings. *dN/dS* rate ratio estimates obtained for the portion of the Antennapedia protein that binds to DNA in a sequence-specific manner were found to be smaller than those obtained for the complete coding sequence or the nhd coding sequence indicating that evolutionary pressure is not homogeneous among different domains within the same protein.

Next the sliding window method was applied to calculate the degree of molecular variation in terms of rates of synonymous and non-synonymous substitutions across the *Antp* gene, moving a window of 60 nucleotides in steps of 20 nucleotides at a time. This window size was chosen because it is approximately one third of the size of the hd nucleotide sequences, which allows thorough sampling of the nucleotide diversity in this region. Results are reported in Figure 4.4, where all measures of *dN/dS* have been plotted against the mid-position of the window along the aligned nucleotide sequence for *Antp*. For this analysis DnaSP uses the substitution model proposed by Nei & Gojobori in 1986 (NG86). In their original paper Nei and Gojobori explained that this simple substitution model gives no weights to different types of codon substitutions. In this model evolutionary rates of synonymous (*dS*) and non-synonymous (*dN*) substitutions are computed by using the two sequences to be compared and transitional and transversional nucleotide substitutions are not considered separately. Thus estimates of synonymous (*dS*) and nonsynonymous (*dN*) rates obtained with this method are slightly different than those obtained using the substitution models implemented by `codeml` from the PAML4 package (see below).

An increase in *dN/dS* values is observed between positions 551 and 698 in Figure 4.4. A *dN/dS* value of 1.128 was observed between Am and Nv in mid-position 638bp. A value of $\omega \geq 1$ indicates that rare, favorable mutations are selected for, resulting in the substitution of the new mutation for the previous best allele (Halliburton, 2004). In pairwise comparisons between species $\omega \geq 1$ indicates positive selection for divergence. In the amino acid alignment this region falls within a run of 8 amino acid residues in Nv, which are not present in Am or in Bt, and one more residue present both in Bt and Nv but not in Am, supporting the theory in favor of diversifying selection acting on this specific region of the gene.

Orange and red boxes (Figure 4.4) denote the positions of the hd as well as a highly conserved run of four amino acid residues YPWM. The YPWM domain is known to be present across the majority of *Hox* genes in Arthropods (Rauskolb and Wieschaus, 1994, Mann and Chan, 1996, Muñoz-Torres and Lawton-Rauh, Data not shown). In the case of *Antp* there is an additional run of 15 conserved amino acid residues located upstream (5') of YPWM. This motif is required for function of most homeotic *Hox* proteins and it interacts with the *Hox* cofactor *extradenticle* (*exd*) in the regulation of expression of downstream genes in higher insects (Johnson et al., 1995). Our data show that in Nv the region between this motif and the homeodomain varies between 13 and 17 amino acids across genes in the cluster (results not shown). The YPWM motif was either lost or modified in the evolution of the 'honorary' *Hox* genes *fushi tarazu* (*ftz*), and *zerknüllt* (*zen*). While the *ftz* homolog in Drosophila lost its YPWM motif, the locust *Schistocerca americana* and the red flour beetle *Tribolium castaneum* homologs maintained it in intact form (Downes et al., 1994 and Brown et al., 1994). Our analyses identified a modified motif **F**PWM in the Nv *ftz* homolog and the presence of the **LXXLL** motif (*Figure C-1.*), which in Drosophila has been shown to interact with Ftz-

F1, a necessary cofactor in Ftz's role in segment formation. All described homologous copies of *zen* in insects have lost this tetrameric motif. This indicates that YPWM was lost somewhere in the evolution of Holometabola.

The *dN/dS* estimates along the hd are within the order of 0 to 0.16 units of evolutionary change, and sequences are highly conserved at the amino acid and nucleotide levels. Figure 4.1 reveals conservation in the *Antp* hd at the nucleotide level among the three species, and Fig 4.4 includes a graphic representation of the degree of conservation at the amino acid level and the amino acid frequency per site with WebLogo (Crooks et al., 2004 and Schneider and Stephens, 1990). Position 898 in the nucleotide alignment, however has an estimated *dN/dS* of 0.328 between Am and Bt, which deviates from the general trend along the homeodomain. In the nucleotide alignment, positions 898-900 are a triplet coding for the amino acid Arginine (R), which may be encoded by six different codons. In this case, the only occurrence within the homeodomain, two different codons are used at this position AGG for Nv and CGC for Am and Bt. Given that position 898 is the midpoint of one of the window in this analysis, it is perhaps this variation what generated an abrupt change in the *dN/dS* rate ratio estimate in this region. It is prudent to remember here that the sliding window graph is a continuum representation of discrete data points, which may occasionally be misleading when interpreting these results. A continuous line between two discrete data points might enhance the effect of the differences between them.

It is evident that peaks from the different pairwise analyses do not always coincide on the same position (e.g. around positions 276-319, 618-638 and 998-1018);

this might be an artifact of the differences in length between the three sequences and the need to choose only one series of ‘mid-point of window’ data points to visualize the graphic comparisons amongst sliding window analyses. In actuality, if it were possible to construct the graph using three X axes simultaneously, all peaks would coincide with each other. Thus, Figure 4.4 is a generalized visual representation of the data amongst the three sequences.

Those regions of the gene where *dN/dS* = 0.0000 are indicated in green rectangles along the length of the nucleotide and amino acid alignment (Figure 4.4). In one end of the spectrum, a value of ω=1 indicates that amino acid changes followed neutral-equilibrium model expectations and they will be fixed at the same rate as a synonymous mutations (Yang and Bielawski, 2000). Estimates of $\omega \leq 1$indicate that amino acid changes are deleterious and that purifying selection will reduce fixation rate. Thus a *dN/dS* estimate of 0.0000 indicates strong purifying selection acting on the DNA sequence to eliminate deleterious mutations and preserve protein function. Hence, green rectangles in Fig 4.4 mark those regions of the gene where purifying selection is acting more strongly, compared to the rest of the protein.

To test if these patterns in *Antp* represent evolutionary processes unique to the *Antp* gene, we compared these results with estimated *dN/dS* values for the antimicrobial peptide *Defensin* (*def*) among *A. mellifera, B. terrestris and N. vitripennis.* Defensins are small cationic peptides, which act primarily against Gram-positive bacteria by electrostatic and hydrophobic interactions leading to disruption of the bacterial membrane (Maget-Dana and Ptak, 1997). Viljakainen and Pamilo (2005) identified and characterized the genomic structure of *Defensin* from the wood ant *Formica aquilonia* and compared its protein structure to that in *A. mellifera* and *Bombus ignitus*. The *Defensin* peptide is made of three domains: a signal peptide, a propeptide and a mature peptide (Fig 4.5). The signal and propeptide domains are proteolitically cleaved to release the active form of the mature peptide (Lazzaro and Clark, 2003). The mature peptide is the portion of the gene that interacts with microbial invaders. As it seemed to be the case with the two different domains in *Antp* (homeobox-encoding and non-homeobox-encoding) we expected to observe a difference in the rates of evolution acting on the different *def* domains. Using SignalP 3.0 we deduced the putative cleavage sites for Bt and Nv *defensin* peptides (Bendsten et al., 2004). Table 4.3 reports the estimated pairwise *dN/dS* rates calculated for the complete *def* coding sequence, and separately for the partial sequences of 'signal & propeptide' and 'mature' domains among the three species following a procedure described by Vijakainen and Pamilo (2008). We found pairwise estimates of evolutionary rates for *def* to be generally slightly higher than those of *Antp* in the same three species by an order of magnitude. With certain exceptions (Lazzaro and Clark, 2003), generally it has been reported that genes involved in the immunity defense mechanisms of animals have faster rates of amino acid substitutions than other nuclear genes (Sackton et al., 2007). Additionally, as it was the case of the *Antp* homeodomain, portions of the *def* peptide which interact with the microbial wall showed lower *dN/dS* estimates than the rest of the protein.

Phylogenetic analyses

Our results suggest that evolutionary changes might be occurring at a different rate along the DNA-binding domain of *Antennapedia*, compared with the rest of the gene. To test for the statistical significance of observed differences in estimated *dN/dS* ratios across the coding sequence of *Antp* we used a phylogenetic approach based on maximum likelihood (ML). Given the differential selective constraints that might be imposed on each codon position (Dónaill and Maktelow, 2004), it is important to consider programs with codon-substitution models that allow partitioning of a dataset according to codon position and consider the codon triplet the unit of evolution (Goldman and Yang, 1994). PAML4 is a package of programs for phylogenetic analyses of DNA and protein sequences that uses ML (Yang, 2007). `codeml` from PAML4 uses ML to estimate the sequence divergence (*t*), the transition/transversion ratio (*κ*) and the ratio *dN*/*dS* (*ω*) from the data using ML, in order to calculate *dN* and *dS* (Yang, 2007). Analyses were conducted for both the full coding sequence of *Antennapedia* (347 codons, excluding gaps) and separately for the homeodomain (59 codons) and nonhomeodomain (288 codons) regions. Results of ML estimates (Table 4.1) confirmed our previous observations of a lower *ω* ratio found along the hd compared with the rest of the *Antp* peptide, and log-likelihood (*ι*) values confirm that *ω* is indeed different from 1 (Yang and Bielawski, 2000).

EMPIRICAL BAYESIAN RECONSTRUCTION OF ANCESTRAL SEQUENCES. *Antennapedia* sequences of extinct ancestors of Apidae (node 5, Figure 4.2) and Aculeta (node 4, Figure

4.2) were reconstructed with `codeml`. Eighty-three synonymous and 14 nonsynonymous changes were observed along branch 1, where the direction of the change goes from the extinct ancestral sequence for the entire group to the extinct ancestral sequence of the bees (node 4 to node 5, Fig 4.1). This likelihood-based reconstruction uses information from branch lengths and relative substitution rates between codons and provides a measure of uncertainties in the form of posterior probabilities for reconstructed ancestral states (Yang, 2008, Yang, 2007). Except for changes observed in positions 172 (p=0.638) and 208 (p=0.559), nonsynonymous changes along branch 1 were estimated with posterior probabilities of 0.995 and higher. Most (13 out of 14) nonsynonymous replacements led to a change in the ionic charge per site (Table 4.4; *Tables C-1.1* through *C-1.4.*).

A summary of the occurrence of all changes along each branch and their posterior probabilities may be found in *Appendix C*. Worthy of notice is that only two nonsynonymous changes were observed in the homeodomain sequence through the reconstructed evolutionary history of the hymenopteran *Antp* gene. One is hypothesized to have occurred along branch 4, which describes changes occurred from the sequence of the extinct ancestor of Aculeta in the direction of Nv (node 3, *Appendix C*). The change from codon AAC (Asparagine) to codon ACC (Threonine) was calculated with a posterior probability of 0.571. The second change occurred in branch 2 from the ancestral *Antp* sequence of the Apidae in the direction of Am at position 286 with a posterior probability of 0.995. In this case the change was from TAC (Tyrosine) to TTC (Phenylananine). Despite the 5 - 37 nonsynonymous changes occurred along each branch

(Table 4.4.), our data suggest that only two such changes have occurred in approximately 195 million years since Chalcidoidea separated from Aculeta (Grimaldi and Engel, 2005). These results provide additional support for our findings of signatures of purifying natural selection along the homeodomain in the Hymenoptera *Antennapedia* gene.

Calculated numbers of synonymous and nonsynonymous changes observed between reconstructed ancestral sequences and extant copies of *Antp* in Hymenoptera are summarized in Table 4.4. It is important to recognize that reconstructed ancestral sequences are not real observed data. Thus, one must practice caution when performing further analysis with them. Using CLUSTAL 2.0.10 we prepared an amino acid multiple sequence alignment including three extant *Antp* sequences and two reconstructed ancestral ones as a visual aid to better understand the natures of the changes here reported (Fig 4.6).

What, then, is the cause of such a low ω rate ratios in *Antp*? Similar analyses with other *Hox* genes and their orthologs could be performed to test the hypothesis that ω estimates are lower in the homeodomain versus non-homeodomain regions in general across loci to see if *Antp* has an overall significantly lower ω estimate versus all other *Hox* genes. As discussed above, in some species of insects the genes zerknüllt (*zen*) and fushi tarazu (*ftz*) have either lost or posses a degenerate YPWM motif. Interestingly neither gene is expressed in a hox-like fashion in insects, i.e. their developmental roles no longer involve a homeotic function. The *ftz* gene plays a role in segmentation of the Drosophila embryo (Carroll and Scott, 1985) and is expressed in a modified pair-rule pattern during development of the short germband red flour beetle *Tribolium castaneum*

(Brown et al., 1994). It is believed that the role of *ftz* in segmentation evolved from that of a canonical Hox with a homeotic function, and that somewhere in between existed an intermediate that played both roles (Hughes and Kaufman, 2002b).

It is now widely known that one duplication event in the ancestral *Hox3* gene of Arthropods gave origin to two copies of *zen* and one of these copies further differentiated into the *bicoid* (*bcd*) gene copy in Drosophila. z*en* plays a role in the development of the extraembryonic ectoderm in Drosophila (Wakimoto et al., 1984) while *bcd* is an anterior-determining morphogen, deposited maternally. In *T. castaneum* the duplication of *zen* was accompanied by subfunctionalization to specify the serosa and later fuse it with the embryonic amnion to complete the dorsal closure during early stages of embryogenesis (van der Zee et al., 2005). These changes in expression patterns, which have modified developmental roles, were also accompanied by sequence changes in which protein functional domains and motifs were added or deteriorated throughout the evolutionary history of these genes (Löhr et al., 2001).

Since *ftz* and *zen* are *Hox* genes which no longer have a homeotic function, perhaps these two genes have different ω estimates when compared to other paralogs in the cluster. As we mentioned above, *Hox3* is ancestral to two copies of z*en*-like genes in insects, with evidence of subfunctionalization occurring after the duplication event in at least one species; *ftz* has also undergone a dramatic change of anatomical domains of gene expression and protein motifs. Thus, if selective constraints are directly impacted in new gene copies such that more non-synonymous mutations accumulate in derived gene copies, then we should see a significant difference between *dN/dS* rate ratios in derived versus ancestral gene copies. Analyses of seven other genes in the *Hox* cluster in Nv suggest an overall degree of conservation across a few structural features besides the homeodomain (*Figure C-1.*). Comparative analyses of the Nv *proboscipedia* (*pb*), *labial* (*lab*), *ftz*, *zen*, *Sex combs reduced* (*Scr*), *Antp*, *Ultrabithorax* (*Ubx*) and *Abdominal-B* (*Abd-B*) showed that other structural features conserved across the eight genes include a run of 10 amino acid residues located 80 to 81 positions upstream of the homeodomain, and conservation of the YPWM motif. The location of the motif varies from gene to gene, as does the spacer region between the motif and the homeodomain. Its sequence is conserved for the most part, with just a single amino acid change in *ftz* and *lab* (*Figure C-1.*). Perhaps performing evolutionary analyses similar to the ones presented here with all other members of the cluster could shed some light on the effects of the apparent relaxation of evolutionary constraint on those paralogs, which have lost their homeotic role subsequent to duplication. Particularly, it would be interesting to explore how the rates of sequence evolution in these genes compare to the ones reported here for *Antp*.

Conclusions

Here we present the newly deduced amino acid sequence for the *Antennapedia* (*Antp*) gene for *Bombus terrestris* and *Nasonia vitripennis* using a BAC shotgun sequencing approach. Combined with currently available data for *Antp* from the genome of *Apis mellifera,* we have used these newly deduced data to study this gene's evolutionary genetic relationships among species of Hymenoptera.

Using maximum likelihood analyses to estimate and compare the substitution rate ratio ω (*dN/dS*) across the *Antp* gene, amongst *Antp* gene regions, and in comparison to another gene (*Defensin*), we detected purifying selection in the overall *Antp* protein sequence. This suggests the paramount importance that this transcription factor plays in the development of the metazoan body plan in Hymenoptera. Additionally, ML analyses performed with `codeml` (PAML4) confirmed our hypothesis that $\omega < 1$ for both the full length and partial coding sequence of *Antp*, indicating that ω estimates are significantly lower along the coding sequence of the homeodomain compared with the rest of the protein. This pattern may be the result of differential evolutionary constraints posed by how selected portions of the protein interact with other proteins or with other regions of the genome.

The fact that *Antp* is expressed and subjected to purifying selection suggests that duplicated gene copies have been maintained possibly by subfunctionalization and/or neofunctionalization (versus pseudofunctionalization and/or non-functionalization). In Arthropods the evolution of novelties in the body plan are believed to be the result of shrinking expression domains along the evolutionary history of the group. More overlapping, larger domains are evident in Chelicerates, with respect to *Hox* expression domains in insects (Averof, 2002). Our results are in agreement with this model because strong purifying selection observed in the *Hox* genes may indicate no functional redundancy among genes, despite similarities in coding sequences. In other words, the genes of the *Hox* Cluster are very conserved at the sequence level, but some sequence variation is maintained, which results in some functional variations that may preserve the genetic framework underlying the development of the Arthropod body plan.

What we have learned so far about the evolutionary history of the genes of the *Hox* cluster, tells us that some members of the ancestral Arthropod cluster have changed so much over evolutionary time that some of them evolved functions which no longer have a homeotic effect and some of them have evolved up to the point where some of them may no longer play a role in development at all, as in the case of one of the copies of *zen* (*zen2*) in Drosophila (Pultz et al., 1988). If selective constraints are directly impacted in new gene copies such that more non-synonymous mutations accumulate in derived gene copies, then we should see a significant difference between *dN/dS* rate ratios in derived versus ancestral gene copies. We would also expect to see a difference in ω estimates for Hox paralogs with non homeotic functions compared those of *Antp* and all other *Hox* which have retained their homeotic capacity.

Finally, we used an empirical Bayesian method of maximum likelihood to reconstruct the *Antennapedia* sequences of extinct ancestral nodes in the evolutionary history of the gene. These sequences are reported here to further illustrate the hypothesized small number of accepted evolutionary changes that have occurred along the *Antp* homeodomain since the Aculeta and Chalcidoidea last shared a common ancestor approximately 195 million years ago.

Experimental Procedures

DNA and amino acid sequences

Sequences for the *Antennapedia* (*Antp*) gene from *Apis mellifera* (Am) were obtained from GenBank (Benson et al., 2008) and were chosen with supporting information from Walldorf and colleagues (2000). GenBank accession numbers for Am *Antp* are AME276511 for the nucleotide sequence and CAC06383 for amino acids. *Antp* sequences for *Bombus terrestris* and *Nasonia vitripennis* (Bt and Nv, respectively) were deduced using a degenerate PCR and BAC Library hybridization approach described in the following sections. Sequences will be deposited in GenBank (accession numbers pending) and are currently found in *Supplementary data C-1*.

Library Screening

Wilfert et al. (2008) and Muñoz-Torres and colleagues (*Unpublished*, *Chapter III*) constructed bacterial Artificial Chromosome (BAC) libraries for Bt and Nv, respectively. Both libraries are available through the Clemson University Genomics Institute (CUGI) at www.genome.clemson.edu. High-density nitrocellulose BAC library filters were screened using a genomic fragment obtained through a degenerate PCR approach. PCR amplification was conducted using primers designed from a multiple sequence alignment following specifications from Rose et al. (1998). *Antp* orthologs were chosen from sequences available at the National Center for Biotechnology Information (NCBI), relying on BLAST (Altschul et al., 1990) sequence-similarity scores with respect to the *Drosophila melanogaster* ortholog to choose the best-fit candidates from other insect species. Selected sequences are listed with species names and GenBank nucleotide and amino acid accession numbers in parenthesis as follows: *Drosophila melanogaster* (AE001572, AAD19793), *Anopheles gambiae* (AF080565, AAC31945), *Tribolium castaneum* (AY043292, AAK96031), *Bombyx mori* (D16684, BAA04087) and *Schistocerca americana* (U32943, AAB03236). Using CLUSTAL X2 (Larkin et al., 2007) we performed a multiple sequence alignment with amino acid sequences from the 5 species. Two conserved motifs (YPRFPPY and VYASCKL) were chosen for primer design. Because we wanted to obtain genomic sequences exclusive of *Antennapedia,* and because both the Homeodomain and the YPWM motifs are highly conserved sequences across arthropods (Hughes and Kaufman, 2002a), conserved residues for primer design were chosen outside these two regions. We used nucleotide sequences corresponding to each conserved motif to design the following degenerate primers. *antp*_FORWARD (5'-3') CCS MGS TTY CCD CCS TAC and *antp*_REVERSE (5'-3') HAR YTT RCA SSW SGY RTA VAC. PCR reactions were performed using species-specific genomic DNA and the Advantage®2 PCR Enzyme System from Clontech (Clontech Laboratories Inc, Mountain View, CA, USA) using the SAPCR buffer and following manufacturer's instructions. Reactions were performed with the following steps 94ºC for 5 min, 94ºC for 30 seconds, 55ºC for 40 seconds, 68ºC for one minute, go to second step 5 times, 94ºC for 30 seconds, 60ºC for 40 seconds, 68ºC for 30 minutes, go to step 6 for 25 cycles, extend at 68ºC for 10 minutes. PCR products were ethanol-purified and a 220bp fragment was gel-isolated using the QIAEX II gel extraction kit (QIAGEN, Valencia, CA, USA). Isolated fragments were cloned onto the pGEM-T Easy vector (Promega Corporation, Madison, WI, USA) and sequenced. Sequenced products were analyzed using BLAST to corroborate that fragments represented indeed a portion of the *Antp* gene from each species. PCR fragments were labeled with 30uCi of radioactive γ-^{32}P using the Random Primed DNA Labeling Kit DECAprime II (Applied Biosystems, Foster City, CA, USA), following manufacturer's specifications. Labeled products were used for hybridization of high density BAC filters using standard techniques (Sambrook et al., 1989) with modifications described in Munoz-Torres et al. (Unpublished, *Chapter III*). Hybridized BAC filters were imaged in a Storm Scanner (GE Healthcare, Piscataway, NJ, USA) and positive hits were scored with HybSweeper (Lazo et al., 2005).

Plasmid DNA from positive BAC hits was prepared from 900 µL cultures of Terrific Broth (GIBCO/Invitrogen Inc. Carlsbad, CA, USA)-Chloramphenicol (12.5 µg/µL) in 96-well format, inoculated with 1.5uL of BAC culture. After 18 hours (h), cultures were treated with a modified alkaline lysis method. To determine insert sizes, approximately 200 ng of BAC DNA from all positive hits were digested using 7 units of *Not*I for 16h at 37°C. DNA digestions were analyzed by PFGE in 1% agarose gels for 15h at 14°C (6 v/cm, switch time of 5-15s) and visualized with Ethidium Bromide. To corroborate the identification of these hits as candidates containing *Antp*, approximately 100ng of BAC DNA was also fingerprinted following standard procedures described in Chen et al., 2002 and Marra et al., 1997 using the library cloning enzyme (*Hind III*) for both BAC libraries. Fingerprinted BAC DNA was transferred onto a positively charged nitrocellulose Hybond membrane (GE Healthcare, Piscataway, NJ, USA) using a Downward Alkaline Transfer Method (Chomezynski, 1992) and transferred fingerprints were hybridized with the same *Antp* probe used on the BAC filters. Once the presence of the *Antp* probe was confirmed, we chose the BAC containing the longest insert for full-length sequencing.

BAC shotgun sequencing

Sub-clone libraries were constructed from candidate BACs from Bt and Nv. BAC DNA was prepared with two replicates per candidate as follows. Individual 3 mL cultures of LB broth (GIBCO/Invitrogen Inc. Carlsbad, CS. USA)-Chloramphenicol (12.5ug/uL) were incubated with a single colony forming unit (cfu) from each candidate BAC and allowed to grow for 4h at 37ºC with agitation at 225rpm. Volume was then raised to 50 mL of LB (Luria broth) and allowed to grow for an additional 16h under equal temperature and shaking conditions. Plasmid DNA was prepared using a standard alkaline lysis method modified for larger volumes; approximately 2 to4µg of DNA were recovered per BAC. DNA was randomly fractured with hydrodynamic shearing forces using a HydroShear (Genomic Solutions, Ann Arbor, MI, USA) with specifications to produce a distribution of sizes with 90% of the DNA falling between 3Kb and 5 Kb according to manufacturer's instructions. DNA was subject to end repair (End-It™ DNA End Repair Kit, Epicentre Biotechnologies, Madison, WI, USA), and posterior size-selection by agarose gel electrophoresis. Fractions were eluted through electrophoresis and ligated into the vector pBluescript II KS+ (Stratagene, La Jolla, CA, USA). Libraries were plated and arrayed into 96-well microtitre plates. Sequencing was performed using the Dye-terminator cycle sequencing kit (Applied Biosystems, Foster City, CA, USA). Sequence data from the forward and reverse priming sites of the shotgun clones were accumulated. Sequence data equivalent to eight (Nv) and six (Bt) times the size of the BAC clones were assembled using `Phred`-`Phrap` programs (Ewing and Green 1998). Sequence editing and assembly confirmation was performed with Consed (Gordon et al., 1998).

Sequence annotation

The assembled consensus sequence from each BAC was analyzed implementing the FGENESH+ (Softberry Inc, Mount Kisco, NY, USA) prediction software combined with BLAST searches on the NCBI website to ascertain the accuracy of the automated prediction. In addition to FGENESH+, the Apollo Genomic Annotation Tool (Lewis et al., 2002) was used to deduce the complete sequence of the Nv *Antp* gene. Decisions on the final version of annotated gene sequences (nucleotide and amino acid) for *Antp* were made based on several factors including sequence similarity compared to other insect homologs, presence of canonical splicing sites in intron/exon boundaries and inspection on the quality of individual reads across the entire nucleotide sequence. Sequences were separately stored in FASTA format for further use.

Molecular Evolution analysis

Multiple Sequence Alignments of the *Antp* gene nucleotide and amino acid sequences were performed with CLUSTAL W 2.0.10 (Larkin et al., 2007). Nucleotide alignments were exported into a PHYLIP format for later use with PAML4. Aligned sequences were also imported into DnaSP (Rozas et al., 2003) to calculate the number of synonymous and non-synonymous substitution differences (*S-Dif* and *N-Dif*, respectively) per total number of synonymous and non-synonymous sites (Table 4.1) using the Nei & Gojobori nucleotide substitution model (NG86, Nei and Gojobori, 1986). Occurrence of non-synonymous sites with multiple (more than 1) changes per codon was detected also with DnaSP and their number and distribution were manually scored on the multiple nucleotide sequence alignment. Data were plotted using MS Excel (Figure 3). Analyses to test for biases in codon usage in any of the species were performed using the 'Codon Usage' command in DnaSP. Estimates of *dN/dS* rates were calculated using the 'DNA Polymorphism and Divergence' command with and without a sliding window option. When the option was implemented, window length was 60 nucleotides with a step size of 20 as reported in Figure 4.4. Putative protein cleavage sites for analysis of *Nasonia defensin* sequences were predicted using SignalP 3.0. (Bendsten et al., 2004). WebLogo was used to measure amino acid frequency per site in the aligned sequences (Crooks et al., 2004 and Schneider and Stephens, 1990) and the results were plotted as shown in Figure 4.4. The logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position for the aligned sequences.

The gene genealogy reconstruction was estimated in MEGA4 (Tamura et al., 2007) using the Neigbor-Joining (NJ) method (Saitou and Nei, 1987) with a K 2-parameter model of sequence evolution (Kimura, 1980). The resulting best-fit tree was exported in Newick format to be used with PAML4.

`codeml` from the PAML4 package was used to measure synonymous and nonsynonymous substitutions rate ratios using a maximum likelihood (ML) algorithm with a codon-based substitution model (Yang, 2007). Log-likelihood values to corroborate pairwise *dN/*dS differences across different portions of the gene were obtained using an F3x4 codon frequency, and assuming 2 or more *dN/dS* ratios for branches. Substitution rates (κ) and ω were also estimated using ML. Estimated ancestral sequence reconstruction was performed using equation 4 (eqn. 4) from Yang et al. (1995) and the `Rate_Ancestor` function from CODONML, part of the `codeml` script in PAML4 (Yang, 2007) with one ω (*dN/dS*) ratio assumed for all branches. The gene genealogy was provided in Newick format (Figure 4.2.). Changes in ionic charge and polarity were putatively identified in a simplistic manner by following the changes in a classification of the 20 amino acids (universal code) according to the character of their side chain or the R group that is bonded to the alpha-carbon in each amino acid available from The Biotechnology Project website (MATC, 2008).

Acknowledgements

Additional genomic sequence data and software training for use with *Nasonia vitripennis* were sponsored by the Nasonia Genome Sequencing Consortium, of which MM-T is part. The authors thank Brad Rauh and Chris Saski for helpful analysis suggestions.

Table 4.1. Maximum likelihood (ML) estimation of evolutionary rates in the *Antennapedia* (*Antp*) gene using a codon-based substitution model for three species of Hymenoptera. Results were obtained using `codeml` from the PAML4 package (Yang, 2007) except in the case of *S-Dif* and *N-Dif*. ML estimates of *t*, *κ*, and *dN/dS* were obtained separately for the complete coding sequence, the Homeodomain and the Non-Homeodomain containing portion of the gene. `codeml` uses these estimates to calculate *dN* and *dS*. *Am* = *Apis mellifera*. *Bt* = *Bombus terrestris*. *Nv* = *Nasonia vitripennis*. *t* = sequence divergence. *S* = total number of synonymous sites. *S-Dif* = total number of synonymous differences. *N* = the total number of nonsynonymous sites. *N-Dif* = total number of nonsynonymous differences. '*' Indicates that this parameter was calculated with DnaSP (Rozas et al., 2003). § Marks the log-likelihood (*ι*) values. (See text for more information).

	t	*S*	*S-Dif**	*N*	*N-Dif**	*dN/dS*(?)	*dN*	*dS*	?§
antp Complete cds									
Am-Bt	0.6047	222.9	105	818.1	23	0.0347	0.0290	0.8349	-1717.6
Am-Nv	1.2811	140.2	118.67	900.8	90.33	0.0401	0.1011	2.5212	-1842.9
Bt-Nv	2.0352	194.1	157	846.9	82	0.0290	0.0936	3.2304	-1918.4
*antp*Non-Homeodomain									
Am-Bt	0.8982	191.8	98	603.2	22	0.0289	0.0329	1.1377	-1318.1
Am-Nv	1.7208	108.4	97.67	686.6	87.33	0.0361	0.1237	3.4238	-1420.9
Bt-Nv	3.2034	161	135	634	80	0.0255	0.1220	4.7908	-1503.9
antp Homeodomain									
Am-Bt	0.1340	22	7	155	1	0.0220	0.0069	0.3115	-255.21
Am-Nv	0.6834	25	21	152	3	0.0089	0.0137	1.5323	-274.42
Bt-Nv	0.7738	25.5	22	151.5	2	0.0039	0.0068	1.7508	-264.13

Table 4.2. Codon bias measurements in three species Hymenoptera. Effective number of codons (ENC), codon bias index (CBI) and Scaled Chi Square (SChi2) were calculated with the Codon Usage command in DnaSP (Rozas et al., 2003) to measure the extent of nonrandom usage of synonymous codons along the sequence of the *Antennapedia* gene in *Apis mellifera* (Am), *Bombus terrestris* (Bt) and *Nasonia vitripennis* (Nv). Reference values are given below. Max – Non = Maximum codon bias – Non-bias. A significant codon bias is evident in the Nv sequence. (See text).

Species	*ENC*	*CBI*	*SChi2*
Am	39.285	0.576	0.509
Bt	50.704	0.293	0.187
Nv	31.904	0.766	0.825
Reference			
Max- Non	20 - 61	1 - 0	1 - 0

Table 4.3. Estimation of synonymous (*dS*) and nonsynonymous (*dN*) substitution rates in 2 different protein-coding regions for three species of Hymenoptera. i) In the *Antennapedia* (*Antp*) gene, *dN*/*dS* rate ratio estimates obtained for the portion of the protein that binds to DNA in a sequence-specific manner are smaller than those obtained for the complete coding sequence and the non-homeodomain coding sequence. ii) In the case of the antimicrobial peptide *Defensin* (def), estimates obtained for *dN*/*dS* are lower over the portion of the protein which becomes the mature peptide in charge of binding to bacterial cells changing their polarity and conformation to lyse them. These results suggest that selective pressure may be acting differently on regions of the protein under different evolutionary constraints, such as specific interactions with other portions of the genome in a sequence-specific manner, or interactions with other cells. Results were obtained using the 'Synonymous and Nonsynonymous substitutions' command in DnaSP. Am = *Apis mellifera*. Bt = *Bombus terrestris*. Nv = *Nasonia vitripennis*.

i)

	antp Complete cds			*antp* Non-Homeodomain			*antp* Homeodomain		
	dS	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*
Am-Bt	0.8459	0.0285	0.0337	1.1425	0.0321	0.0281	0.3126	0.0069	0.0220
Am-Nv	2.7455	0.1044	0.0380	3.9107	0.1283	0.0328	1.5321	0.0137	0.0089
Bt-Nv	3.2704	0.0901	0.0276	4.8298	0.1172	0.0243	1.7505	0.0068	0.0039

ii)

	def Complete cds			*def* Signal & Propeptide			*def* Mature		
	dS	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*	*dS*	*dN*	*dN/dS*
Am-Bt	0.9370	0.1513	0.1615	1.1329	0.1854	0.1637	0.8204	0.1272	0.1550
Am-Nv	2.8306	0.4230	0.1494	2.5838	0.6705	0.2595	3.0761	0.2810	0.0913
Bt-Nv	1.2804	0.3550	0.2773	1.7761	0.5593	0.3149	1.0421	0.2354	0.2259

Table 4.4. Number of nonsynonymous (n) and synonymous (s) changes observed in modern copies of the *Antennapedia* gene respect to their reconstructed ancestral sequences. Sequences from extinct ancestors of *A. mellifera* and *B. terrestris* and of this group (Apidae) and the *N. vitripennis* (Chalcidoidea) were reconstructed using the `Rate_Ancestor` feature of `codeml` from the PAML4 package (Yang, 2007). 'Polarity' and 'Charge' indicate the total number of nonsynonymous changes, which led to a change of this kind, per branch. All changes in polarity also led to changes in charge, but not all changes in charge were necessarily accompanied by changes in polarity.

Branch	*node to node*	*n*	*s*	*Polarity*	*Charge*	*Total*
1	4 to 5	14	83	10	13	97
2	5 to 1	14	40	2	7	54
3	5 to 2	5	64	3	4	69
4	4 to 3	37	51	17	25	88
Total	-	70	238	-	-	308

Figure 4.1. Nucleotide sequence alignment of *Antennapedia* (*Antp*) gene complete-CDS from three species of Hymenoptera constructed using CLUSTAL 2.0.10 (Larkin et al., 2007). The fourth track indicates the degree of nucleotide sequence conservation; '*' denotes that nucleotides in that column are identical in all sequences. Pairwise alignment scores are as follows: sequences (1:2), score: 80; sequences (1:3), score: 77; sequences (2:3), score: 87. Sequence 1 corresponds to *Nasonia vitripennis* (nv), 2 to *Apis mellifera* (am), and 3 is its homolog in *Bombus terrestris* (bt). The homeobox (orange) and the conserved YPWM motif (red) are indicated with colored boxes.

```
am_antp_CDS_1059bp     ATGAGTTCGTATTTCGCGAATTCGTACATCCCGGACCTGCGTAATGGCGGGGTGGAACAC 60
bt_antp_CDS_1059bp     ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGACCTGCGTAATGGCGGGGTGGAACAC 60
nv_antp_CDS_1089bp     ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGATCTGCGCAATGGCGGGGTGGAGCAT 60
                       *********** *********************** ***** ************** **

am_antp_CDS_1059bp     CCGCATCAGCATCAGCAGCACTACGGTGCGGCCGTCCAGGTGCCCCAGCAGACGCAGTCG 120
bt_antp_CDS_1059bp     CCGCATCAGCATCAGCAGCACTACGGTGCCGCCGTCCAGGTGCCCCAGCAAACGCAGTCG 120
nv_antp_CDS_1089bp     CCGCATCAGCACCAGCAGCACTACGGCGCGGCGGTCCAGGTGCCCCAGCAGCAGCAGGCC 120
                       *********** ************** ** ** *****************   **** *

am_antp_CDS_1059bp     GTGCAGCAACAGTCCCAGCAGGCCGGGGACCCGTGCGACCCGAGCCTGCTACGCCAGGGC 180
bt_antp_CDS_1059bp     GTACAGCAGCCATCTCAGCAAACCGGGGATCCATGCGATCCTAGTCTCCTACGTCAGGGA 180
nv_antp_CDS_1089bp     GTGCAGCAGCAGCCCCAGCAGGCGAGCGACCCCTGCGACCCGTCGATGCTGCGCCAAGGC 180
                       ** ***** *   * *****  *  * ** ** ***** **     * ** ** ** **

am_antp_CDS_1059bp     GTGCCCGGCCACCAT---TACGGGGCGGCGGGCAGCCAGCAA---GACATGCCTTATCCG 234
bt_antp_CDS_1059bp     GTGCCTGGCCATCAC---TATGGGGCCGCTGGTAGCCAGCAA---GATATGCCTTATCCG 234
nv_antp_CDS_1089bp     GTGCCGGGCCACCACGGCTACGGGGCCGCGACGGGCCAGCAGCCGGGCATGCCCTACCCC 240
                       ***** ***** **    ** ***** **     *******    *  ***** ** **

am_antp_CDS_1059bp     AGGTTCCCGCCCTACAACCGGATGGACATGCGTAACGCGACGTATTATCAGCACCAACAG 294
bt_antp_CDS_1059bp     AGGTTTCCTCCGTACAATCGGATGGACATGCGGAACGCGACCTATTATCAGCATCAACAG 294
nv_antp_CDS_1089bp     CGCTTCCCGCCCTACGACCGCATGGACATCAGGAACGCGGCCTACTACCAGCAGCAGCAG 300
                        * ** ** ** *** * ** ********  * ****** * ** ** ***** ** ***

am_antp_CDS_1059bp     GACCACGGGAGCGGGATGGACGGGATGGGTGGTTACAGGTCGGCGTCGCCGAGCCCTGGC 354
bt_antp_CDS_1059bp     GAGCACGGCAGC---ATGGACGGGTTGGGTGGTTACAGGTCGACGTCCCCGAGCCCCGGT 351
nv_antp_CDS_1089bp     CAGGAGCACGGC---ATGGAC---ATGGCCAGCTACCGGGCGAGCTCGCCGAGCGCGGGC 354
                        *  *     **   ******    ***   * *** ** **   ** ****** * **

am_antp_CDS_1059bp     ATGGGC------CACATGGGGCACACGCCGACCCCT---AACGGGCACCCG---TCCACC 402
bt_antp_CDS_1059bp     ATGGGC------CACATGGGACACACACCGACCCCG---AACGGACATCCG---TCCACT 399
nv_antp_CDS_1089bp     ATGGCCGGCCTCCACATGGGCCACACGCCGACCCCGGTCAACGGCCACCCCGCCAGCACG 414
                       **** *      ******** ***** ********    ***** ** **  ***   *

am_antp_CDS_1059bp     CCGATCGTGTACGCGAGCTGCAAGCTGCAGGCGGCGGCGGTCGATCACCAGGGGAGCGTG 462
bt_antp_CDS_1059bp     CCTATTGTCTATGCGAGTTGCAAGCTTCAAGCGGCCGCGGTCGATCATCAGGGTAGCGTA 459
nv_antp_CDS_1089bp     CCCATCGTCTACGCGAGCTGCAAGCTCCAAGCGGCGGCGGTCGACCACCAGGGCAGCGTC 474
                       ** ** ** ** ***** ******** ** ***** ******** ** ***** *****

am_antp_CDS_1059bp     CTCGACGGGCCGGACAGCCCGCCGCTGGTCGAGTCGCAGATGCACCACCAAATGCACACG 522
bt_antp_CDS_1059bp     CTCGATGGACCGGACAGCCCGCCATTGGTCGAGTCGCAGATGCACCACCAAATGCATTCG 519
nv_antp_CDS_1089bp     CTCGACGGGCCCGATAGTCCGCCGCTCGTCGACGCCCAGATGCACCACCAGATGCACCCC 534
                       ***** ** ** ** ** *****  * *****  * ************** *****  *

am_antp_CDS_1059bp     CAACACCCCCACATGCAGCCGCAGCAGGGCCAGCACCAGTCG------------------ 564
bt_antp_CDS_1059bp     CAACATCCTCACATGCAGCCGCAACAGTCACAACATCAACAGCAG--------------- 564
nv_antp_CDS_1089bp     CAGCACACGCACATGCAGGCCCAGCAGTCGCACCCCCAGCAGCAGCCCCAGCCTCAAGCG 594
                       ** **  * ********* * ** ***   ** *  **   *
```

```
am_antp_CDS_1059bp      ---------CAAGCACAGCAGCAGCATCTTCAGGCGCACGAGCAGCACATGATGTACCAG 615
bt_antp_CDS_1059bp      ---------CAGCAGCAGCATCAACATCTTCAGGCGCAGCAGCAGCACATGATGTACCAA 615
nv_antp_CDS_1089bp      CCTCACCAGCAGGCCCACATGCAACCCCAGCAGACGCAGCAGCAGCACATGATGTACCAG 654
                                 **    **    ** *     *** ****  *******************

am_antp_CDS_1059bp      CAGCAGCAGCAGTCGCAGGCTGCCTCGCAGCAGTCGCAGCCAGGCATGCACCCGCGACAG 675
bt_antp_CDS_1059bp      CAGCAACAACAAACACAGGCGGCGTCGCAACAATCTCAGCCTGGCATGCATCCGCAACAA 675
nv_antp_CDS_1089bp      CAGCAGACGCAGCCCCAG---------CAGCCCCAGCCCGCGGCGATGCACCCCCAGCAG 705
                        *****    **  * ***         ** *     *   * *  ***** ** *  **

am_antp_CDS_1059bp      CAGCAGCAAGCTCAGCAACACCAAGGGGTGGTCACGTCGCCGCTAAGCCAGCAGCAACAG 735
bt_antp_CDS_1059bp      CAACAGCAACCTCAGCAACACCAAGGGGTGGTCACGTCGCCGCTTAGTCAGCAACAGCAG 735
nv_antp_CDS_1089bp      CAGGCCCAGCAGCAGCAGCACCAGGGCGTCGTCGCCTCGCCGCTCGGCCAGCAGCAGCCC 765
                        **    **    ***** ***** ** ** *** * ********  * ***** ** *

am_antp_CDS_1059bp      GCCGCGCCTCAGGGCGCGGCAAGCGCCAACCTACCGAGCCCTCTGTACCCGTGGATGAGA 795
bt_antp_CDS_1059bp      GCCGCTCCTCAAGGTGCGGCCACTGCCAACCTACCAAGTCCGCTCTACCCGTGGATGAGA 795
nv_antp_CDS_1089bp      GGCACGCCCCAGAGCGCCGCGCCGACGAACCTGCCCAGTCCCCTCTACCCCTGGATGAGG 825
                        *** * ** **  * ** **     * ***** ** ** ** ** ***** ********

am_antp_CDS_1059bp      AGTCAATTCGAGAGGAAACGAGGCCGGCAAACGTACACCCGATACCAAACCCTCGAGCTC 855
bt_antp_CDS_1059bp      AGTCAATTCGAGAGGAAACGAGGCCGGCAAACGTACACCCGATACCAAACCCTCGAGCTG 855
nv_antp_CDS_1089bp      AGTCAGTTTGAGAGGAAGCGTGGCCGGCAGACGTACACGCGATACCAGACCCTCGAGCTC 885
                        ***** ** ******** ** ******** ******** ******** ***********

am_antp_CDS_1059bp      GAGAAGGAGTTCCACTACAACCGATACCTGACCAGGCGGCGTCGCATCGAGATCGCGCAC 915
bt_antp_CDS_1059bp      GAGAAGGAGTTCCACTTCAACCGATACCTGACCAGGCGGCGGCGCATCGAGATCGCGCAC 915
nv_antp_CDS_1089bp      GAGAAGGAGTTCCACTTCAACCGCTACCTGACCAGGCGACGCAGGATCGAGATCGCGCAT 945
                        **************** ****** ************** **  * ***************

am_antp_CDS_1059bp      GCCCTCTGCCTTACCGAGCGGCAAATCAAAATCTGGTTTCAAAACAGACGGATGAAATGG 975
bt_antp_CDS_1059bp      GCACTCTGCCTGACGGAACGGCAAATCAAAATCTGGTTCCAAAACAGACGGATGAAATGG 975
nv_antp_CDS_1089bp      GCGCTCTGCCTGACCGAGCGCCAGATCAAGATCTGGTTCCAGAACAGGCGCATGAAGTGG 1005
                        ** ******** ** ***** ** ***** ******** ** ***** ** ** ** ***

am_antp_CDS_1059bp      AAGAAGGAGAACAAGTCGAAGGGCACGCCCGGCTCGGGCGACGGGGACACCGAGATCTCG 1035
bt_antp_CDS_1059bp      AAGAAGGAGAACAAGACGAAGGGCGAACCGGGCTCGGGCGACGGCGACACTGAAATCTCG 1035
nv_antp_CDS_1089bp      AAAAAGGAGACCAAGACGAAGGGCGAGCCGAACTCGGGAGACGGTGACACCGACATCTCG 1065
                        ** ************ ********  **   ****** ***** ***** ** ******

am_antp_CDS_1059bp      CCGCAGACGTCGCCGCAGGGTTGA 1059
bt_antp_CDS_1059bp      CCGCAGACATCGCCGCAGGGTTGA 1059
nv_antp_CDS_1089bp      CCCCAGACCTCGCCCCAGGGTTGA 1089
                        ** ***** ***************
```

Figure 4.2. Gene genealogy of *Antennapedia* (*Antp*) from *A. mellifera* (am), *Bombus terrestris* (bt) and *Nasonia vitripennis* (nv). Reconstructed using the Neighbor-Joining (NJ) method (Saitou and Nei, 1987) in MEGA (Kumar et al., 2008). Diamonds represent ancestral species at nodes (n) 4 and 5. The optimal tree with the sum of branch length = 0.37038142 is shown. The scale bar represents estimates of evolutionary distances calculated with the Kimura 2-parameter model (Kimura, 1980) as number of base substitutions per site and are as follows: ((amantp:0.04661,btantp:0.0884):0.0490,nvantp:0.1864).

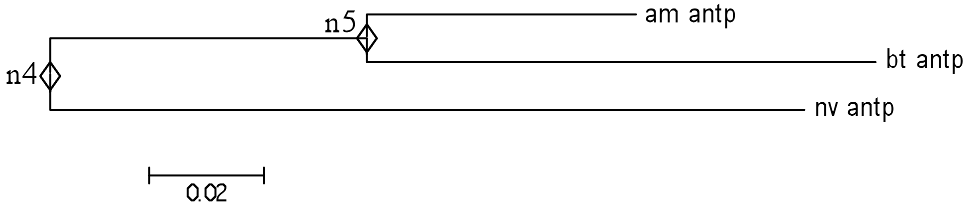
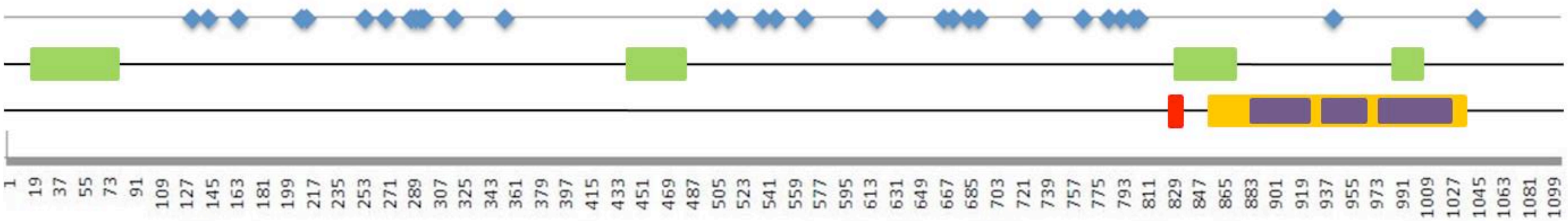


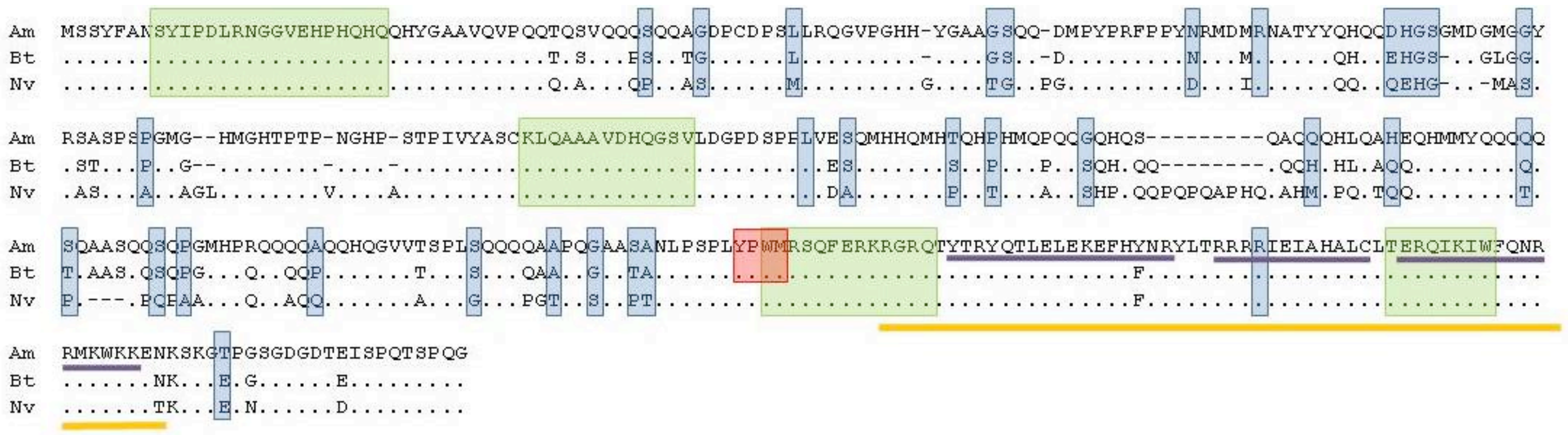
Fig. 4.3 Structural features of the *Antennapedia* (*Antp*) gene in three species of Hymenoptera. i) The nucleotide sequence alignment of *Antp* from *Apis mellifera* (am), *Bombus terrestris* (Bt) and *Nasonia vitripennis* (Nv) is 1,104 bp long (alignment not shown). Blue diamonds mark the occurrence of codons with more than one nucleotide change. Note the very low occurrence of such changes in or around the Homeodomain (orange box) compared to the rest of the protein. Green rectangles denote regions of this alignment for which *dN/dS* values are 0 when measuring this rates ratio with the sliding window method (see Materials and Methods), indicating that no conservative nor semi-conservative substitutions occur among the three species in these regions. The red box indicates the location of a conserved tetramer of amino acid residues before the Homeodomain, characteristic of most Hox genes (see text). Purple rectangles inside the Homeodomain mark the location of each of the helices of the helix-turn-helix-turn-helix motif encoded by the Homeodomain. ii) Amino acid alignment of *Antp* gene. Synonymous substitutions are marked with a ‘.’ and semi-conservative substitutions and alignment gaps are shown. Colored boxes correspond with the previous description, except the Homeodomain and the helices are underlined, rather than boxed.

i)



1
19
37
55
73
91
109
127
145
163
181
199
217
235
253
271
289
307
325
343
361
379
397
415
433
451
469
487
505
523
541
559
577
595
613
631
649
667
685
703
721
739
757
775
793
811
829
847
865
883
901
919
937
955
973
991
1009
1027
1045
1063
1081
1099

ii)



Am MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQSVQQQSQQAGDPCDPSLLRQGVPGHH-YGAAGSQQ-DMPYPRFPPYNRMDMRNATYYQHQQDHGSGMDGMGGY
Bt .....................................T.S...PS..TG......L.........-....GS..-D.........N...M......QH..EHGS-..GLGG.
Nv .....................................Q.A...QP..AS......M.........G....TG..PG.........D...I......QQ..QEHG-..-MAS.
Am RSASPSPGMG--HMGHTPTP-NGHP-STPIVYASCKLQAAAVDHQGSVLDGPDSPFLVESQMHHQMHTQHPHMQPQQGQHQS---------QAQQQHLQAHEQHMMYQQQQQ
Bt .ST...P..G--........-....-................................ES.......S..P...P..SQH.QQ---------.QQH.HL.AQQ.........Q.
Nv .AS...A..AGL........V....A................................DA.......P..T...A..SHP.QQPQPQAPHQ.AHM.PQ.TQQ.........T.
Am SQAASQQSQPGMHPRQQQQAQQHQGVVTSPLSQQQQAAPQGAASANLPSPLYPWMRSQFERKRGRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHALCLTERQIKIWFQNR
Bt T.AAS.QSQPG...Q..QQP.......T...S...QAA..G..TA.............................F..
Nv P.---.PQPAA...Q..AQQ.......A...G...PGT..S..PT.............................F..
Am RMKWKKENKSKGTPGSGDGDTEISPQTSPQG
Bt .......NK...E.G......E.........
Nv .......TK...E.N......D.........

Fig. 4.4. Evolutionary studies of *Antennapedia* in Hymenoptera species. i) Sliding window analyses of the ratio of rates of evolutionary change (*dN/dS*) of three *Antennapedia* sequences from *Apis mellifera* (Am), *Bombus terrestris* (Bt) and *Nasonia vitripennis* (Nv). Analyses were performed using DnaSP (Rozas et al, 2003) as described in the text. This graph was constructed using an amino acid-based corrected nucleotide alignment. (See text for more information). ii) Amino acid frequency per site in the aligned sequences was measured and plotted with WebLogo (Crooks et al., 2004 and Schneider and Stephens, 1990). Amino acids are presented as stacked letters at each position on the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.

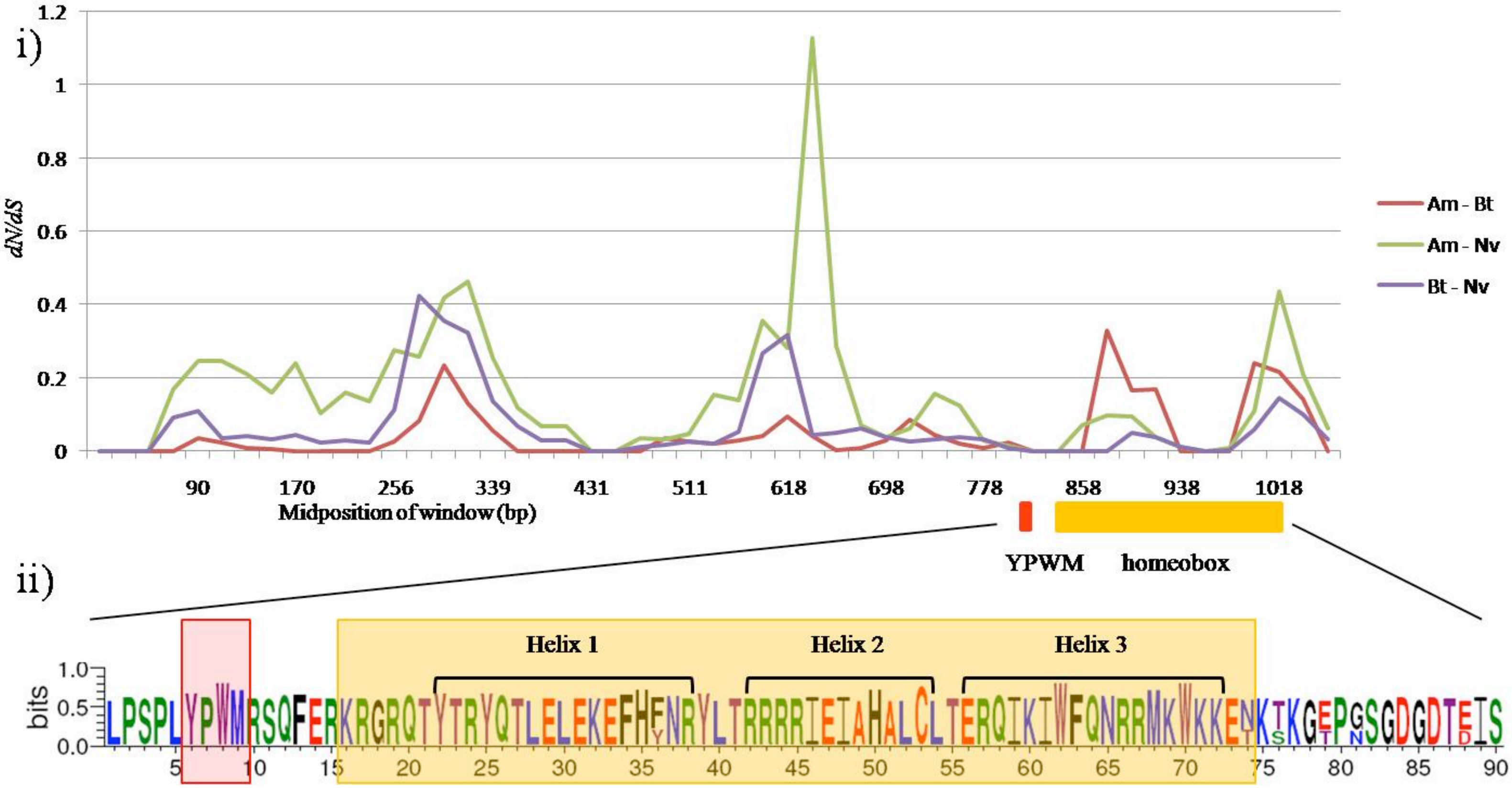
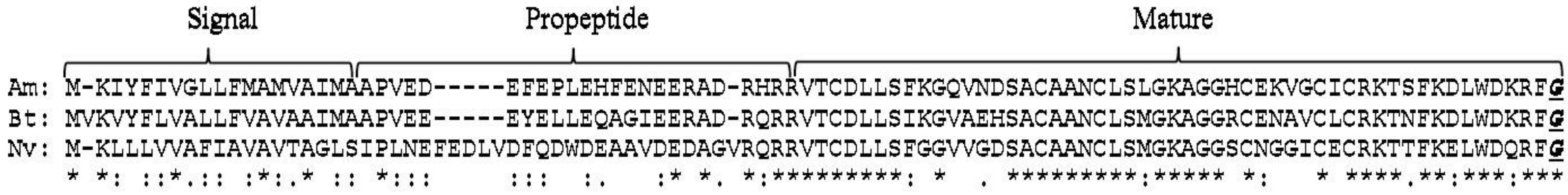

i)
1.2
1
0.8
0.6
0.4
0.2
0
dN/dS
90
170
256
339
431
511
618
698
778
858
938
1018
Midposition of window (bp)
Am - Bt
Am - Nv
Bt - Nv
YPWM
homeobox
ii)
Helix 1
Helix 2
Helix 3
bits
1.0
0.5
0.0
LPSPLYPWMRSQFERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSGDGDTEIS
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90

Fig. 4.5. Amino acid alignment of *defensin* sequences from *Apis mellifera* (Am), *Bombus terrestris* (Bt) and *Nasonia vitripennis* (Nv). Brackets above the alignment mark 3 differentially cleaved protein domains. A fourth track under the aligned amino acids marks the degree of conservation across the sequences as follows: ‘*’ denotes that residues in that column are identical in all sequences in the alignment. ‘:’ means that conserved substitutions have been observed. ‘.’ indicates that semi-conserved substitutions are observed. The last Glycine residue at the Carboxyl end of each sequence is a donor of amidation and does not make part of the mature peptide (Casteels-Josson et al, 1994). GenBank Accession Numbers are as follows: Am: NM_001011616, Bt: FJ161700 and Nv: XM_001603271. Multiple amino acid sequence alignment was performed with CLUSTAL 2.0.10 (Larkin et al, 2007) and cleavage sites were predicted using SignalP 3.0 (Bendsten et al, 2004).



```
Am: M-KIYFIVGLLFMAMVAIMAAPVED-----EFEPLEHFENEERAD-RHRRVTCDLLSFKGQVNDSACAANCLSLGKAGGHCEKVGCICRKTSFKDLWDKRFG
Bt: MVKVYFLVALLFVAVAAIMAAPVEE-----EYELLEQAGIEERAD-RQRRVTCDLLSIKGVAEHSACAANCLSMGKAGGRCENAVCLCRKTNFKDLWDKRFG
Nv: M-KLLLVVAFIAVAVTAGLSIPLNEFEDLVDFQDWDEAAVDEDAGVRQRRVTCDLLSFGGVVGDSACAANCLSMGKAGGSCNGGICECRKTTFKELWDQRFG
    * *: ::*.:: :*:.* :: *:::     :::  :.   :* *. *:*********: *   .*********:***** *:   * ****.**:***:***
```

Figure 4.6. Amino acid sequence alignment of *Antennapedia* (*Antp*) gene complete-CDS from three species of Hymenoptera and their reconstructed ancestral sequences. Nodes 4 and 5 in the phylogenetic tree reconstructed in Figure 1 are considered the ancestral species to the extant hymenopterans here analyzed. Node 5 is the extinct ancestor of Apidae and node 4 the one for Aculeta

```
CLUSTAL 2.0.10 multiple sequence alignment

antp_am         MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQSVQQQSQQAGDPCDPSLLRQG 60
antp_bt         MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQSVQQPSQQTGDPCDPSLLRQG 60
Node_5          MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQSVQQQSQQAGDPCDPSLLRQG 60
Node_4          MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQAVQQQSQQAGDPCDPTLLRQG 60
antp_nv         MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQQQAVQQQPQQASDPCDPSMLRQG 60
                ************************************* *:*** .**:.*****::****

antp_am         VPGHH-YGAAGSQQ-DMPYPRFPPYNRMDMRNATYYQHQQDHGSGMDGMGGYRSASPSPG 118
antp_bt         VPGHH-YGAAGSQQ-DMPYPRFPPYNRMDMRNATYYQHQQEHGS-MDGLGGYRSTSPSPG 117
Node_5          VPGHH-YGAAGSQQ-DMPYPRFPPYNRMDMRNATYYQHQQEHGS-MD-MGGYRSTSPSPG 116
Node_4          VPGHH-YGAAGSQQ-DMPYPRFPPYNRMDMRNATYYQHQQEQDS-MD-MGGYRSTSPSPG 116
antp_nv         VPGHHGYGAATGQQPGMPYPRFPPYDRMDIRNAAYYQQQQQEHG-MD-MASYRASSPSAG 118
                ***** **** .** .*********:***:***:***:**:. . ** :..**::***.*

antp_am         MG--HMGHTPTP-NGHP-STPIVYASCKLQAAAVDHQGSVLDGPDSPPLVESQMHHQMHT 174
antp_bt         MG--HMGHTPTP-NGHP-STPIVYASCKLQAAAVDHQGSVLDGPDSPPLVESQMHHQMHS 173
Node_5          MG--HMGHTPTP-NGHP-STPIVYASCKLQAAAVDHQGSVLDGPDSPPLVESQMHHQMHS 172
Node_4          MG--HMGHTPTP-NGHP-STPIVYASCKLQAAAVDHQGSVLDGPDSPPLVEAQMHHQMHP 172
antp_nv         MAGLHMGHTPTPVNGHPASTPIVYASCKLQAAAVDHQGSVLDGPDSPPLVDAQMHHQMHP 178
                *.  ******** **** ********************************::*******.

antp_am         QHPHMQPQQGQHQS---------QAQQQHLQAHEQHMMYQQQQQSQAASQQSQPGMHPRQ 225
antp_bt         QHPHMQPQQSQHQQQ--------QQQHQHLQAQQQHMMYQQQQQTQAASQQSQPGMHPQQ 225
Node_5          QHPHMQPQQSQHQQ---------QAQQQHLQAQQQHMMYQQQQQTQ---QQSQPGMHPQQ 220
Node_4          QHPHMQPQQSQHQQ---------QAQQQHLQAQQQHMMYQQQQQPQ---QPPPPGMHPQQ 220
antp_nv         QHTHMQAQQSHPQQQPQPQAPHQQAHMQPQQTQQQHMMYQQQTQPQ---QPQPAAMHPQQ 235
                **.***.**.: *.         * : *  *:::******** *.*   *   ..***:*

antp_am         QQQAQQHQGVVTSPLSQQQQAAPQGAASANLPSPLYPWMRSQFERKRGRQTYTRYQTLEL 285
antp_bt         QQQPQQHQGVVTSPLSQQQQAAPQGAATANLPSPLYPWMRSQFERKRGRQTYTRYQTLEL 285
Node_5          QQQPQQHQGVVTSPLSQQQQAAPQGAATANLPSPLYPWMRSQFERKRGRQTYTRYQTLEL 280
Node_4          QPQPQQHQGVVASPLSQQQPAAPQGAATANLPSPLYPWMRSQFERKRGRQTYTRYQTLEL 280
antp_nv         QAQQQQHQGVVASPLGQQQPGTPQSAAPTNLPSPLYPWMRSQFERKRGRQTYTRYQTLEL 295
                * * *******:***.*** .:**.**.:*******************************

antp_am         EKEFHYNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKSKGTPGSGDGDTEIS 345
antp_bt         EKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSGDGDTEIS 345
Node_5          EKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSGDGDTEIS 340
Node_4          EKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPDSGDGDTEIS 340
antp_nv         EKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKETKTKGEPNSGDGDTDIS 355
                *****:*************************************.*:** *.******:**

antp_am         PQTSPQG 352
antp_bt         PQTSPQG 352
Node_5          PQTSPQG 347
Node_4          PQTSPQG 347
antp_nv         PQTSPQG 362
                *******
```

References

Akashi, H, A Eyre-Walker, 1998. Translational selection and molecular evolution. *Curr Opin Genet Dev* **8**:688-693.

Allendorf, FW, GH Thorgaard. 1984. Tetraploidy and the evolution of salmonid fishes. In: Turner BJ, editor. Evolutionary genetics of fishes. New York: Plenum Press. p1-46.

Altschul, SF, W Gish, W Miller, EW Myers & DJ Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403-410.

Angelini, DR and TC Kaufman. 2005.Comparative Developmental Genetics and the Evolution of Arthropod Body Plans. *Annu Rev Genet* **39**:95-119.

Averof, M. 2002. Arthropod *Hox* genes: insights on the evolutionary forces that shape gene functions. *Curr Op Genet Dev* **12**:386-392.

Bachiller, D, A Macias, D Douboule, and G Morata. 1994. Conservation of a functional hierarchy between mammalian and insect Hox/HOM genes. *EMBO J* **13**:1930-1941.

Bendtsen JD, H Nielsen, G von Heijne and S Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**:783-795.

Benson, DA, I Karsch-Mizrachi, DJ Lipman, J Ostell, DL Wheeler. 2008. GenBank. *Nucleic Acids Res* **36**(Database issue):D25-30.

Brown, SJ, RB Hilgenfeld, and RE Denell. 1994. The beetle *Tribolium castaneum* has a *fushi tarazu* homolog expressed in stripes during segmentation. *Proc Natl Acad Sci USA* **91**:12922-12926.

Casteels-Josson, K, W Zhang, T Capaci, P Casteels and P Tempst. 1994. Acute transcriptional response of the honeybee peptide-antibiotics gene repertoire and required post-translational conversion of the precursor structures. *J Biol Chem* **269**: 8569-28575.

Carroll, SB. 2000. Endless Forms: The Evolution of Gene Regulation and Morphological Diversity. *Cell* **101**:577-580.

Carroll, SB and MP Scott. 1985. Localization of the *fushi tarazu* protein during Drosophila embryogenesis. *Cell*. **43**(1):47-57.

Chantawannakul, P and RW Cutler. 2008. Convergent host-parasite codon usage between honey bee and bee associated viral parasites. *J Inv Path* **98**(2):206-210.

Chen, M, G Presting, WB Barbazuk, JL Goicoechea, B Blackmon, G Fang, H Kim, D Frisch, Y Yu, S Sun, S Higingbottom, J Phimphilai, D Phimphilai, S Thurmond, B Gaudette, P Li, J Liu, J Hatfield, D Main, K Farrar, C Henderson, L Barnett, R Costa, B Williams, S Walser, M Atkins, C Hall, MA Budiman, JP Tomkins, M Luo, I Bancroft, J Salse, F Regad, T Mohapatra, NK Singh, AK Tyagi, C Soderlund, RA Dean, RA Wing. 2002. An Integrated Physical and Genetic Map of the Rice Genome. *Plant Cell* **14:**537-545.

Chomezynski, P. 1992. One hour downward alkaline capillary transfer for blotting of DNA and RNA. *Analytical Biochemistry* **201**:134-139.

Coen, ES and EM Meyerowitz. 1991. The war of the whorls: Genetic interactions controlling flower development. *Nature* **353**:31-37.

Crooks, GE, G Hon, JM Chandonia, SE Brenner. 2004.WebLogo: A sequence logo generator. *Genome Research* **14**:1188-1190.

Dawes, R, I Dawson, F Falciani, G Tear, and M Akam. 1994. *Dax*, a locust *Hox* gene related to *fushi-tarazu* but showing no pair-rule expression. *Development* **120**:1561-1572.

Denell, R. 1994. Discovery and genetic definition of the Drosophila *Antennapedia* complex. *Genetics* 138:549-552.

Deutsch, JS, E Mouchel-Vielh. 2003. Hox genes and the crustacean body plan. *Bioessays* **25**(9):878-87.

Dónaill, DAM and M Manktelow. 2004. Molecular Informatics: Quantifying Information Patterns in the Genetic Code. *Molecular Simulation* **30**(5):267-272.

Ewing, B, L Hillier, M Wendl, P Green. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research* **8**:175-185.

Gehring, WJ. 1987. Homeo boxes in the study of development. *Science* **236**:1245-1252.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725-736.

Gordon, D, C Abajian, and P Green. 1998. Consed: A Graphical Tool for Sequence Finishing. *Genome Research* **8**:195-202.

Grimaldi, DA and MS Engel. 2005. Evolution of the Insects. Cambridge University Press. New York, NY. USA. 755p.

Halliburton, R. 2004. Introduction to population genetics. Pearson Prentice Hall. Upper Saddle River, NJ. USA. 650p.

Hayward, DC, NH Patel, EJ Rehm, CS Goodman, and EE Ball. 1995. Sequence and expression of grasshopper *Antennapedia*: comparison to Drosophila. *Dev Biol* **172**:452-465.

Heuer, JG, and TC Kaufman. 1992. Homeotic genes have specific functional roles in the establishment of the Drosophila embryonic peripheral nervous system. *Development* **115**: 35-47.

Hoegg S, A Meyer. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet* **21**(8):421-424.

Hughes, CL and TC Kaufman. 2002a. *Hox* genes and the evolution of the arthropod body plan. *Evol Dev* **4**(6):459-499.

Hughes, CL, and TC Kaufman. 2002b. Exploring the myriapod body plan: expression patterns of the ten *Hox* genes in a centipede. *Development* **129**:1225-1238.

Johnson, FB, E Parker, MA Krasnow. 1995. *Extradenticle* protein is a selective cofactor for the Drosophila homeotics: Role of the homeodomain and YPWM amino acid motif in the interaction. *Proc Natl Acad Sci USA* **92**(3):739-743.

Jukes, TH and CR Cantor. 1969. Evolution of protein molecules. In HN Munro (ed.), Mammalian Protein Metabolism. Academic Press, New York. p21-132.

Kaufman, TC, MA Seeger and G Olsen. 1990. Molecular and genetic organization of the *Antennapedia* gene complex of *Drosophila melanogaster*. *Adv Gen* **27**:309-362.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J of Mol Evol* **16**:111-120.

Kumar, S, J Dudley, M Nei and K Tamura. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* **9**:299-306.

Larkin, MA, G Blackshields, NP Brown, R Chenna, PA McGettigan, H McWilliam, F Valentin, IM Wallace, A Wilm, R Lopez, JD Thompson, TJ Gibson and DG Higgins. 2007. ClustalW and ClustalX version 2. *Bioinformatics* **23**(21):2947-2948.

Lazzaro, BP and AG Clark. 2003. Molecular Population Genetics of Inducible Antibacterial Peptide Genes in *Drosophila melanogaster*. *Mol Biol Evol* **20**(6):914-923.

Lazo, GR, N Lui, YQ Gu, XY Kong, D Coleman-Derr and OD Anderson. 2005. Hybsweeper: a resource for detecting high-density plate gridding coordinates. *Biotechniques* **39**:320, 322, 324.

Lemmons, D and W McGinnis. 2006. Genomic Evolution of *Hox* Gene Clusters. *Science* **313**:1918-1922.

Lewis, EB.1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**:565-570.

Lewis, RA, TC Kaufman, RE Denell, P Tallerico.1980a. Genetic Analysis of the *Antennapedia* Gene Complex (Ant-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. I. Polytene Chromosome Segments 84b-D *Genetics* **95**(2):367-381

Lewis, RA, BT Wakimoto, RE Denell, TC Kaufman. 1980b. Genetic Analysis of the *Antennapedia* Gene Complex (Ant-C) and Adjacent Chromosomal Regions of *Drosophila melanogaster*. II. Polytene Chromosome Segments 84A-84B1, 2. *Genetics* **95**(2):383-397.

Lewis, SE, SMJ Searle, N Harris, M Gibson, V Iyer, J Richter, C Wiel, L Bayraktaroglu, E Birney, MA Crosby, JS Kaminker, BB Matthews, SE Prochnik, CD Smith, JL Tupy, GM Rubin, S Misra, CJ Mungall and ME Clamp. 2002. Apollo: a sequence annotation editor. *Genome Biology* **3**(12):research0082.1-0082.14.

Löhr, U, M Yussa and L Pick. 2001. Drosophila *fushi tarazu*. a gene on the border of homeotic function. *CB* **11**:1403-1412.

Lynch, VJ, JJ Roth and GP Wagner. 2006. Adaptive evolution of *Hox*-gene homeodomains after cluster duplications. *BMC Evol Biol* **6**:86-98.

Madison Area Technical College (MATC). 2008. Chapter 2: Protein Structure. Laboratory Manual.
http://matcmadison.edu/biotech/resources/proteins/labManual/chapter_2.htm

Maget-Dana, R and M Ptak. 1997. Penetration of the insect defensin A into phospholipid monolayers and formation of defensin A-lipid complexes. *Biophys J* **73**(5):2527-2533.

Mann, RS, and SK Chan. 1996. Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. *Trends Gen* **12**:258-262.

Marchler-Bauer, A, JB Anderson, MK Derbyshire, C DeWeese-Scott, NR Gonzales, M Gwadz, L Hao, S He, DI Hurwitz, JD Jackson, Z Ke, D Krylov, CJ. Lanczycki, CA Liebert, C Liu, F Lu, S Lu, GH Marchler, M Mullokandov, JS Song, N Thanki, RA Yamashita, JJ Yin, D Zhang, and S H Bryant. 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35**(D):237-240.

Marra, M, K Dewar, P Dunn, JR Ecker, S Fischer, S Kloska, H Lehrach, M Marra, R Martienssen, S Meier-Ewert and T Altmann. 1997. High throughput fingerprint analysis of large-insert clones: Contig construction and selection of clones for DNA-sequencing. *Genome Res* **7**:1072-1084.

McGinnis, W, RL Garber, J Wirz, A Kuroiwa and WJ Gehring. 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37**:403-408.

Morton, BR. 1993. Chloroplast DNA codon use: Evidence for selection at the *psb A* locus based on tRNA availability. *J Mol Evol* **37**:273-280.

Nagy, A, H Hegyi, K Farkas, H Tordai, E Kozma, L Bányai and L Patthy. 2008. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics* **9**:353-378.

Newton, MA, B Mau, B Larget. 1998. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Statistics and Molecular Biology. 30p.

Pultz, MA, RJ Diederich, DL Cribbs, TC Kaufman. 1988. The *proboscipedia* locus of the *Antennapedia* complex: a molecular and genetic analysis. *Genes Dev* **2**:901-920.

Rauskolb, C, E Wieschaus. 1994. Coordinate regulation of downstream genes by extradenticle and the homeotic selector proteins. *EMBO J* **13**:3561-3569.

Reuter, R, and MP Scott. 1990. Expression and function of the homeotic genes *Antennapedia* and *Sex combs reduced* in the embryonic midgut of Drosophila. *Development* **109**:289-304.

de Rosa, R, JK Grenier, T Andreeva, CE Cook, A Adoutte, M Akam, SB Carroll, Balavoine G.1999. Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**:772-776.

Rose, TM, ER Schultz, JG Henikoff, S Pietrokovski, CM McCallum and S Henikoff. 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* **26**(7):1628-1635.

Rozas, J, JC Sánchez-DelBarrio, X Messeguer and R Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**:2496-2497.

Saitou, N. and M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**(4):406-425.

Sackton, TB, BP Lazzaro, TA Schlenke, JD Evans, D Hultmark and AG Clark. 2007. Dynamic evolution of the innate immune system in Drosophila. *Nature Genetics* **39**:1461-1468.

Sambrook, J, E Fritsch, and T Maniatis. 1989. Molecular cloning: A laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 1659p.

Schneider, TD, RM Stephens. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res* **18**:6097-6100.

Shields, DC, PM Sharp, DG Higgins and F Wright. 1988. "Silent"sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons. *Mol Bio Evol* **5**:704-716.

Salamov, A and V Solovyev. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10:**516-522.

Struhl, G. 1982. Genes controlling segmental specification in the Drosophila thorax. *Proc Natl Acad Sci USA* **79:**7380-7384.

Tamura, K, J Dudley, M Nei & S Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**:1596-1599.

Viljakainen L and P Pamilo. 2005. Identification and molecular characterization of defensin gene from the ant *Formica aquilonia*. *Insect Mol Biol* **14**(4):335-338.

van der Zee, M. 2005. Distinct functions of the *Tribolium zerknüllt* genes in serosa specification and dorsal closure. *Curr Biol* **15**(7):624-36.

Wakimoto, BT, FR Turner and TC Kaufman. 1984. Defects in embryogenesis in mutants associated with the *antennapedia* complex of *Drosophila melanogaster*. *Dev Biol* **102**: 147-172.

Walldorf, U, P Binner and R Fleig. 2000. Hox genes in the honey bee *Apis mellifera*. *Dev Genes Evol* **210**(10):483-492.

Watterson, GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* **7**:256-276.

Wilfert, L, M Muñoz Torres, C Reber-Funk, R Schmid-Hempel, J Tomkins, J Gadau and P Schmid-Hempel. 2008. Construction and characterization of a BAC-library for a key pollinator, the bumblebee *Bombus terrestris* L. DOI: 10.1007/s00040-008-1034-1.

Wolfe, KH and WH Li. 2003. Molecular evolution meets the genomics revolution. *Nat Genet* **33:**Suppl, 255-265.

Wright, F. 1990. The "effective number of codons" used in a gene. *Gene* 87:23-29.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586-1591.

Yang, Z. 2008. PAML: Phylogenetic Analysis by Maximum Likelihood. User Guide. 63p.

Yang, ZH and JP Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15** (12):496-503.

Yang, Z, S Kumar, and M Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641-1650.

CHAPTER FIVE

FUTURE PROSPECTS

As for the development of genomic resources, comparative research into the social Hymenoptera stands to gain much by combining information from the sequenced genome of the honey bee (The Honey Bee Genome Consortium, 2006) and the jewel wasp (Nasonia Genome Sequencing Consortium, Unpublished), as well as genomic information from related species. Research communities will greatly benefit from orchestrating a concerted effort leading to sequencing the genome of their organisms of interest. For instance, other traditional genetic approaches to gene finding have been exhausted, and when obtaining the whole genome sequence is still in process or has not yet started, genomic libraries represent the most effective tool for interrogating the genome.

Tools such as a BAC library represent a starting point for the generation of genomic data in species of interest. For example, with the help of a honey bee BAC library, the complementary sex determination gene (*csd*) was identified and demonstrated to be functional. Similarly, the sex determination locus in *B. terrestris* has been genetically mapped to an approximate location. Using information from the available genome projects in species of Hymenoptera, and the three BAC libraries described here, it will be possible to rapidly investigate the molecular and genetic nature of many important features in these species. The paramount importance of these species in agricultural and public health affairs ensures the participation of a strong scientific community that will very likely use these resources exhaustively.

After performing molecular evolution analyses on one of the *Hox* genes from three species of Hymenoptera, we were able to conclude that not only is the synonymous and nonsynonymous substitutions rate ratio (*dN/dS*) low for *Antennapedia*, (*Antp*), but also that *dN/dS* estimates are different between the homeodomain and non-homeodomain protein regions. What lies behind and beyond the estimation of these values? If selective constraints are directly impacted in new gene copies throughout evolutionary history such that more non-synonymous mutations accumulate in derived gene copies, then we should see a significant difference between *dN/dS* rate ratios in derived versus ancestral gene copies. We would also expect to see a difference in ω estimates for *Hox* paralogs with non-homeotic functions compared those of *Antp* and all other *Hox* which have retained their homeotic capacity.

Evolutionary analyses similar those performed here with *Antp* on all other members of the cluster could shed some light on the effects of the apparent relaxation of evolutionary constraints on paralogs, which have lost their homeotic role. Estimates of the rates of evolution in all other genes compared to the ones reported here for *Antp* will very likely expand our understanding of the processes that shaped the evolution of the insect *Hox* cluster of genes, and thus improve our understanding of mechanisms responsible for large and small scale developmental differences in insects.

The fact that *Antp* is expressed and subjected to purifying selection suggests that duplicated genes have been maintained possibly by subfunctionalization and/or neofunctionalization (versus pseudofunctionalization and/or non-functionalization). Our results show that strong purifying selection is observed within the *Hox* genes, which may indicate no functional redundancy among genes, despite similarities in coding sequences. In other words, the genes of the *Hox* Cluster are very conserved at the sequence level, but some level of variation is maintained that translates into functional variations and presumably preserves the genetic framework underlying the development of the Arthropod body plan.

In vertebrates, researchers have identified ultra-conserved regions outside the homeodomain across different orders. These are coding regions that are 120-nucleotide long in *Hox* genes, and are believed to have an important role in their expression and functions. In a similar manner, our analyses of all *Nasonia Hox* genes identified a conserved region of 30 nucleotides, located upstream from the homeodomain, in the coding region. Functional analyses with this conserved domain will help to test whether such region also plays a functional role in the observed patterns of *Hox* genes in *Nasonia*.

Lastly, events of gene transfers between organisms of different species provide a different perspective in the study of the origin and evolution of genes. Such studies become more complex when the two organisms in question belong to distant branches of the evolutionary tree. As reported in *Appendix A*, research reporting on widespread lateral gene transfer (LGT) from intracellular bacteria to multicellular eukaryotes showed that some of the inserted bacterial genes are transcribed within eukaryotic cells; this suggests that these heritable lateral gene transfers may provide a mechanism for acquisition of new genes and functions. What are these mechanisms? How do they drive origins and evolutionary diversification of genes? How have both eukaryotic machinery and genes (and the genomic vicinity) changed since the bacterial insertion event? The next step in this study could be to address these questions and others related to the transmission of the inserted DNA and how endosymbionts, located in the germlines of their eukaryotic hosts, manipulate host cell biology and reproduction. Additionally, Appendix B contains data concerning ongoing research to address similar questions related to a lateral gene transfer event reported from endosymbiotic *Wolbachia pipientis* into the genome of the tropical fruit fly *Drosophila ananassae*. Prospective goals stemming from this portion of this dissertation will be to study the sequence evolution and selection across insert variants in the Hawai'i and Mali *D. ananassae* strains. Studies of sequence variation along the insert are important to investigate selection acting on insert DNA and to elucidate what has happened to the LGTs since insertion. How is selection acting on the inserted DNA? Do endosymbiotic bacterial accelerate divergence and speciation of their eukaryotic host? Future studies conducted to answer these questions will also provide valuable information describing the patterns of evolution after the LGT event and shed light on the mechanisms underlying a wide range of genome processes ultimately responsible for evolutionary diversification of genes and genomes.

APPENDICES

Appendix A

Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes.

From Dunning Hotopp, JC, ME Clark, DCSG Oliveira, JM Foster, P Fischer, MC Muñoz Torres, JD Giebel, N Kumar, N Ishmael, S Wang, J Ingram, RV Nene, J Shepard, J Tomkins, S Richards, DJ Spiro, E Ghedin, BE Slatko, H Tettelin, JH Werren. 2007. Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science* **317** (5845):1753-1756. Reprinted with permission from AAAS. License number: 2107690668137. License Date: Jan 14, 2009.

Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes

Julie C. Dunning Hotopp[1,*], Michael E. Clark[2,*], Deodoro C. S. G. Oliveira[2], Jeremy M. Foster[3], Peter Fischer[4], Mónica C. Muñoz Torres[5], Jonathan D. Giebel[2], Nikhil Kumar[1], Nadeeza Ishmael[1], Shiliang Wang[1], Jessica Ingram[3], Rahul V. Nene[1,‡], Jessica Shepard[1,§], Jeffrey Tomkins[5], Stephen Richards[6], David J. Spiro[1], Elodie Ghedin[1,7], Barton E. Slatko[3], Hervé Tettelin[1,†], John H. Werren[2,†]

[1] The Institute for Genomic Research, J. Craig Venter Institute, 9712 Medical Center Dr., Rockville, MD 20850, USA.

[2] Department of Biology, University of Rochester, Rochester, NY 14627, USA.

[3] Molecular Parasitology Division, New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, USA.

[4] Department of Internal Medicine, Infectious Diseases Division, Washington University School of Medicine, St. Louis, MO 63110, USA.

[5] Clemson University Genomics Institute, 304 BRC, 51 New Cherry St, Clemson, SC 29634, USA.

[6] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

[7] Division of Infectious Diseases, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA.

[*,†] These authors contributed equally to this work.

Correspondence should be addressed to JCDH. E-mail: jdunning@jcvi.org

[‡]Present address: Brown University, Providence, RI 02912, USA.

[§]Present address: Pace University, New York, NY 10038, USA.

ABSTRACT

Although common among bacteria, lateral gene transfer— the movement of genes between distantly related organisms— is thought to occur only rarely between bacteria and multicellular eukaryotes. However, the presence of endosymbionts within some eukaryotic germlines, such as that observed in *Wolbachia pipientis*, may facilitate bacterial gene transfers to eukaryotic host genomes. We therefore examined host genomes for evidence of gene transfer events from intracellular bacteria of the genus *Wolbachia* to their hosts. We found transfers into the genomes of 4 insect and 4 nematode species that ranged from nearly the entire *Wolbachia* genome (>1 Mb) to short (<500 bp) insertions. We also show that some of these inserted *Wolbachia* genes are transcribed within eukaryotic cells lacking endosymbionts. Therefore, heritable lateral gene transfer occurs into eukaryotic hosts from their prokaryote symbionts, potentially providing a mechanism for acquisition of new genes and functions.

The transfer of DNA between diverse organisms, lateral gene transfer (LGT), facilitates the acquisition of novel gene functions. Among Eubacteria, LGT is involved in the evolution of antibiotic resistance, pathogenicity, and metabolic pathways (*1*). Rare LGT events have also been identified between higher eukaryotes with segregated germ cells(*2*), demonstrating that even these organisms can acquire novel DNA. Although most described LGT events occur within a single domain of life, LGT has been described both between Eubacteria and Archaea (*3*) and between prokaryotes and phagotrophic unicellular eukaryotes (*4, 5*). However, few interdomain transfers involving higher multicellular eukaryotes have been found.

Wolbachia pipientis is a maternally inherited endosymbiont that infects a wide range of arthropods, including at least 20% of insect species, as well as filarial nematodes (*6*). It is present in developing gametes (*6*) and so provides circumstances conducive for heritable transfer of bacterial genes to the eukaryotic hosts. *Wolbachia*-host transfer has been described in the bean beetle *Callosobruchus chinensis* (*7*) and in the filarial nematode *Onchocerca* spp. (*8*).

We have found *Wolbachia* inserts in the genomes of additional diverse invertebrate taxa, including fruit flies, wasps, and nematodes. A comparison of the published genome of the *Wolbachia* endosymbiont of *Drosophila melanogaster* (*9*) and assemblies of *Wolbachia* clone mates (*10*) from fruit fly whole genome shotgun sequencing data revealed a large *Wolbachia* insert in the genome of the widespread tropical fruit fly *Drosophila ananassae*. Numerous contigs were found that harbored junctions between *Drosophila* retrotransposons and *Wolbachia* genes. The large number of these junctions and the deep sequencing coverage across the junctions indicated that these inserts were probably not due to chimeric libraries or assemblies. To validate these observations, five *Drosophila*-*Wolbachia* junctions were PCR amplified and three end-sequence verified. Fluorescence *in situ* hybridization (FISH) of banded polytene chromosomes with fluorescein-labeled probes of two *Wolbachia* genes (*11*) revealed the presence of *Wolbachia* genes on the 2L chromosome of *D. ananassae* (Fig. 1).

We found that nearly the entire *Wolbachia* genome was transferred to the fly nuclear genome as evidenced by the presence of PCR amplified products from 44/45 physically distant *Wolbachia* genes from cured strains of *D. ananassae* Hawaii (those verified by microscopy to be lacking the endosymbiont after treatment with antibiotics) (*11*). In contrast, only spurious, incorrectly-sized, and weak amplification was detected from a cured control line lacking these inserts (Townsville). The 45 genes assayed (Table S1) are spaced throughout the *Wolbachia* genome. Thus the high proportion of amplified genes suggests gene transfer of nearly the entire *Wolbachia* genome to the insect genome.

A14 kB region containing four *Wolbachia* genes with two retrotransposon insertions was sequenced (*11*) from a single BAC**,** constituting an independent source of DNA as compared to the largely plasmid-derived whole genome sequence of *D. ananassae*. The two retroelements each contained 5 bp target site duplications (9/10 bp identical), long terminal repeats, and *gag*-*pol* genes (Fig. 2A) indicating that the *Wolbachia* insert is accumulating retroelements. Insertion of this region appears to be recent, as shown by the nearly identical target site duplications and >90% nucleotide identity between the corresponding endosymbiont genes and the sequenced homologs in the *D. ananassae* chromosome.

Crosses between *Wolbachia*-free Hawaii males (with the insert) and *Wolbachia*-free Mexico females (without the insert) revealed that the insert is paternally inherited by offspring of both sexes, confirming that *Wolbachia* genes are inserted into an autosome. Since *Wolbachia* infections are maternally inherited this also confirms that PCR amplification in the antibiotic treated line is not due to a low level infection. Furthermore, the Hawaii and Mexico crosses revealed Mendelian, autosomal inheritance of *Wolbachia* inserts (paternal N=57, *k*=0.49; maternal N=40, *k*= 0.58). Six physically distant, inserted *Wolbachia* genes perfectly co-segregated in F2 maternal inheritance crosses (*11*), suggesting they also are closely linked.

PCR amplification and sequencing (*11*) of 45 *Wolbachia* loci in 14 *D. ananassae* lines from widely dispersed geographic locations revealed large *Wolbachia* inserts in lines from Hawaii, Malaysia, Indonesia, and India (Table S2). Sequence comparisons of the amplicons from these four lines revealed that all ORFs remained intact with >99.9% identity between inserts. This is compared to an average 97.7% identity for the inserts compared to *w*Mel, the *Wolbachia* endosymbiont of *D. melanogaster*. These results indicate the widespread prevalence of *D. ananassae* strains with similar inserts of the *Wolbachia* genome, probably due to a single insertion from a common ancestor.

In addition, RT-PCR followed by sequencing (*11*) demonstrated that ~2% of *Wolbachia* genes (28 of 1206 genes assayed; Table S3) are transcribed in cured adult males and females of *D. ananassae* Hawaii. The complete 5' sequence of one of the transcripts, WD_0336, was obtained using 5'-RACE on uninfected flies (*11*) suggesting that this transcript has a 5' mRNA cap, a form of eukaryotic post-transcriptional modification. Analysis of the transcript levels of inserted *Wolbachia* genes with qRT-PCR (*11*) revealed that they are 10^4-fold to10^7-fold less abundant than the fly's highly transcribed Actin gene (*act5C*; Table S3). There is no cutoff that defines a biologically relevant level of transcription, and assessment of transcription in whole insects can obscure important tissue specific transcription. Therefore, it is unclear whether these transcripts are biologically meaningful, and further work is needed to determine their significance.

Screening of public shotgun sequencing data sets has identified several additional cases of LGT in different invertebrate species. In *Wolbachia* cured strains of the wasp *Nasonia,* six small *Wolbachia* inserts (<500 bp) were verified by PCR and sequencing (*11*) that are >96% nucleotide identity to native *Wolbachia* sequences, in some cases with short insertion site duplications. These include four in *Nasonia vitripennis;* one in *Nasonia giraulti;* and one in *Nasonia longicornis* (Table S4; Fig. 2B). Amplification and sequencing of 14-18 geographically diverse strains of each species indicated that the inserts are species-specific. For example, three *Wolbachia* inserts in *N. vitripennis* are not found in the closely related species *N. giraulti* or *N. longicornis*, which diversified ~1 million years ago (*12*). These data suggest that the *Wolbachia* gene inserts are of relatively recent origin, similar to the inserts in *D. ananassae*.

Nematode genomes also contain inserted *Wolbachia* sequences. As *Wolbachia* infection is required for fertility and development of the worm *Brugia malayi*, the genomes of both organisms were sequenced simultaneously complicating assemblies and leading to the removal of *Wolbachia* reads during genome assembly [>98% identity over 90% of the read length on the basis of the independent BAC-based genome sequence of *w*Bm, the *Wolbachia* endosymbiont of *B. malayi* (*13*)]. Despite this, the genome of *B. malayi* contains 249 contigs with *Wolbachia* sequences (e-value<10^{-40}); nine of which were confirmed by long range PCR and end sequencing (*11*). These include eight large scaffolds containing >1 kb *Wolbachia* fragments within 8 kb of a *B. malayi* gene (Table S5). Comparisons of *w*Bm homologs to these regions suggested that all of these *Wolbachia* genes within the *B. malayi* genome are degenerate. In addition, a single region <1 kb was examined that contains a degenerate fragment of the *Wolbachia* aspartate aminotransferase gene (Wbm0002). Its location was confirmed by PCR and sequencing in *B. malayi* as well as *B. timori* and *B. pahangi* (*11*).

Of the remaining 21 arthropod and nematode genomes in the trace repositories (*11*), we found six containing *Wolbachia* sequences. Potential host-*Wolbachia* LGT was detected in three: *Drosophila sechellia, Drosophila simulans,* and *Culex pipiens* (Table 1).

The sequencing of *w*Bm also facilitated the discovery of a *Wolbachia* insertion in *Dirofilaria immitis* (dog heartworm). The *D. immitis Dg2* chromosomal region encoding the D34 immunodominant antigen (*14, 15*) contains *Wolbachia* DNA within its introns and in the 5'-UTR (Fig. 2C). These *Wolbachia* genomic fragments have maintained synteny with the *w*Bm genome (*13*), suggesting they may have inserted as a single unit and regions were replaced by exons of *Dg2*. A second chromosomal region (*DgK*) has been identified in other *D. immitis* lines that has 91% nucleotide identity in the exon sequences but contains differing number, position, size, and sequence of introns (*16*) and has no homology to known *Wolbachia* sequences.

Whole eukaryote genome sequencing projects routinely exclude bacterial sequences on the assumption that these represent contamination. For example, the publicly available assembly of *D. ananassae* does not include any of the *Wolbachia* sequences described here. Therefore, the argument that the lack of bacterial genes in these assembled genomes indicates that bacterial LGT does not occur is circular and invalid. Recent bacterial LGT to eukaryotic genomes will continue to be difficult to detect if bacterial sequences are routinely excluded from assemblies without experimental verification. And these LGT events will remain understudied despite their potential to provide novel gene functions and impact arthropod and nematode genome evolution. Because *W. pipientis* is among the most abundant intracellular bacteria (*17, 18*), and its hosts are among the most abundant animal phyla, the view that prokaryote to eukaryote transfers are uncommon and unimportant needs to be reevaluated.

References and Notes

1. Y. Boucher *et al.*, *Annu Rev Genet* **37**, 283 (2003).

2. S. B. Daniels, K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, A. Chovnick, *Genetics* **124**, 339 (1990).

3. K. E. Nelson *et al.*, *Nature* **399**, 323 (1999).

4. J. O. Andersson, *Cell Mol Life Sci* **62**, 1182 (2005).

5. W. F. Doolittle, *Trends Genet* **14**, 307 (1998).

6. R. Stouthamer, J. A. Breeuwer, G. D. Hurst, *Annu Rev Microbiol* **53**, 71 (1999).

7. N. Kondo, N. Nikoh, N. Ijichi, M. Shimada, T. Fukatsu, *Proc Natl Acad Sci U S A* **99**, 14280 (2002).

8. K. Fenn *et al.*, *PLoS Pathog* **2**, e94 (2006).

9. M. Wu *et al.*, *PLoS Biol* **2**, e69 (2004).

10. S. L. Salzberg *et al.*, *Genome Biol* **6**, R23 (2005).

11. Information on materials and methods is available on Science Online.

12. B. C. Campbell, J. D. Steffen-Campbell, J. H. Werren, *Insect Mol Biol* **2**, 225 (1993).

13. J. Foster *et al.*, *PLoS Biol* **3**, e121 (2005).

14. S. Sun, K. Sugane, *J Helminthol* **68**, 259 (1994).

15. S. H. Sun, T. Matsuura, K. Sugane, *J Helminthol* **65**, 149 (1991).

16. K. Sugane, K. Nakayama, H. Kato, *J Helminthol* **73**, 265 (1999).

17. J. H. Werren, *Annu Rev Entomol* **42**, 587 (1997).

18. J. H. Werren, D. M. Windsor, *Proc Biol Sci* **267**, 1277 (2000).

19. The *D. ananassae* BAC 01L18 sequence is deposited in Genbank (EF426679), the *D. ananassae* sequence comparisons from the 4 lines are deposited in Genbank (EF611872-EF611985), the *D. ananassae* RACE sequence is deposited in dbEST (46867557) and Genbank (ES659088), the *Nasonia* sequences are deposited in Genbank (EF588824-EF588901), the *Brugia malayi* contigs are available in Genbank (DS237653, DS238272, DS238705, DS239028, DS239057, DS239291, DS239377, DS239315), the *Brugia malayi* scaffolds are available in Genbank

(AAQA01000958, AAQA01000097, AAQA01001500, AAQA01000425, AAQA01001819, AAQA01001498, AAQA01000384, AAQA01001952, AAQA01000736, AAQA01000571, AAQA01000369), and the microarray primers sequences that were used for RT-PCR are deposited in ArrayExpress as (A-TIGR-28). This work was supported by an NSF grant to JHW and HT, an NIAID grant to EG, and New England Biolabs Inc. support to BES.

Supporting Online Material

www.sciencemag.org

Materials and Methods

Tables S1 to S5

References

Table A-1. Summary of *Wolbachia* sequences and evidence for LGT in public databases.

				Junctions validated	
	Total Traces	*Wolbachia*		*by PCR and sequencing*	*Wolbachia infection*
Organism	*Screened*	*traces*	*LGT*	*(successful/attempted)*	*described in literature*
*Trace Repository Sequences**					
Acyrthosiphon pisum (aphid)	4,285,120	0			+
Aedes aegypti (mosquito)	16238263	0			-
Anopheles gambiae (mosquito)	5,456,630	0			-
Apis mellifera (honeybee)	3,941,137	0			-
Brugia malayi (filarial nematode)	1,260,214	22524	+	12-Oct	+
Culex pipiens quinquefasciatus (mosquito)	7,380,430	21304	+	0/0	+
Daphnia pulex (crustacean)	2,724,768	0			-
Drosophila ananassae (fruit fly)	3,878,537	38605	+	7-Jun	+
Drosophila erecta (fruit fly)	2,916,936	0			-
Drosophila grimshawi (fruit fly)	2,874,111	0			-
Drosophila melanogaster (fruit fly)	1,001,855	0			+
Drosophila mojavensis (fruit fly)	3,130,180	107†	-		-
Drosophila persimilis (fruit fly)	1,375,313	0			-
Drosophila pseudoobscura (fruit fly)	5,161,792	0			-
Drosophila sechellia (fruit fly)	1,203,722	1	+	0/0	+
Drosophila simulans (fruit fly)	2,321,958	7473	+	0/0	+
Drosohila virilis (fruit fly)	3,632,492	0			-
Drosophila willistoni (fruit fly)	2,332,565	2519	-		+
Drosophila yakuba (fruit fly)	2,269,952	0			+
Ixodes scapularis (tick)	13,088,763	44	-		+
Nasonia giraulti (wasp)	540,102	2	+	1-Jan	+
Nasonia longicornis (wasp)	447,736	1	+	1-Jan	+
Nasonia vitripennis (wasp)	3,360,694	30	+	4-Apr	+
Pediculus humanus (head louse)	1,480,551	0			+
Pristonchus pacificus (nematode)	2,292,543	0			-
Tribolium castaneum (beetle)	1,918,906	0			-
Genbank sequences					
Dirofilaria immitis (filarial nematode)	NA	NA	+		+

* All whole genome shotgun sequencing reads were downloaded for 26 arthropod and nematode genomes (*11*). Organisms identified as lacking *Wolbachia* sequences either had no match or matches only to the prokaryotic rRNA. Since the *Nasonia* genomes are from antibiotic-cured insects, they were identified as having a putative LGT event merely on identification of *Wolbachia* sequences in a read. All other organisms were considered to have putative LGT events if the trace repository contained ≥1 read with (a) >80% nucleotide identity over 10% of the read to a characterized eukaryotic gene, (b) >80% identity over 10% of the read to a *Wolbachia* gene, and (c) manual review of the BLAST results for 1-20 reads to ensure significance (*11*).

† This isolate was previously shown to have *Wolbachia* reads in its trace repositories that are contaminating reads from the *D. ananassae* genome sequencing project (*10*).

Figure A-1. Fluorescence microscopy evidence supporting *Wolbachia/*host LGT. DNA in the polytene chromosomes of *D. ananassae* are stained with propidium iodide (red) while a probe for the *Wolbachia* gene WD_0484 binds to a unique location (green, arrow) on chromosome 2L.

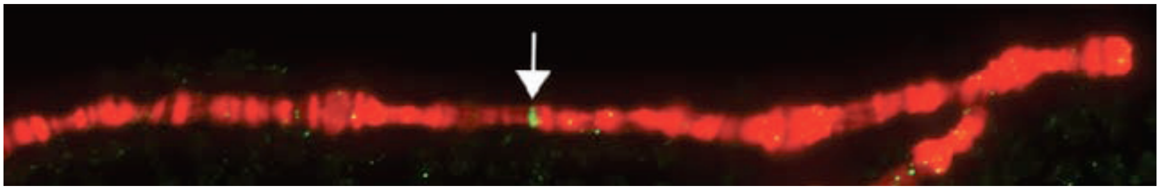
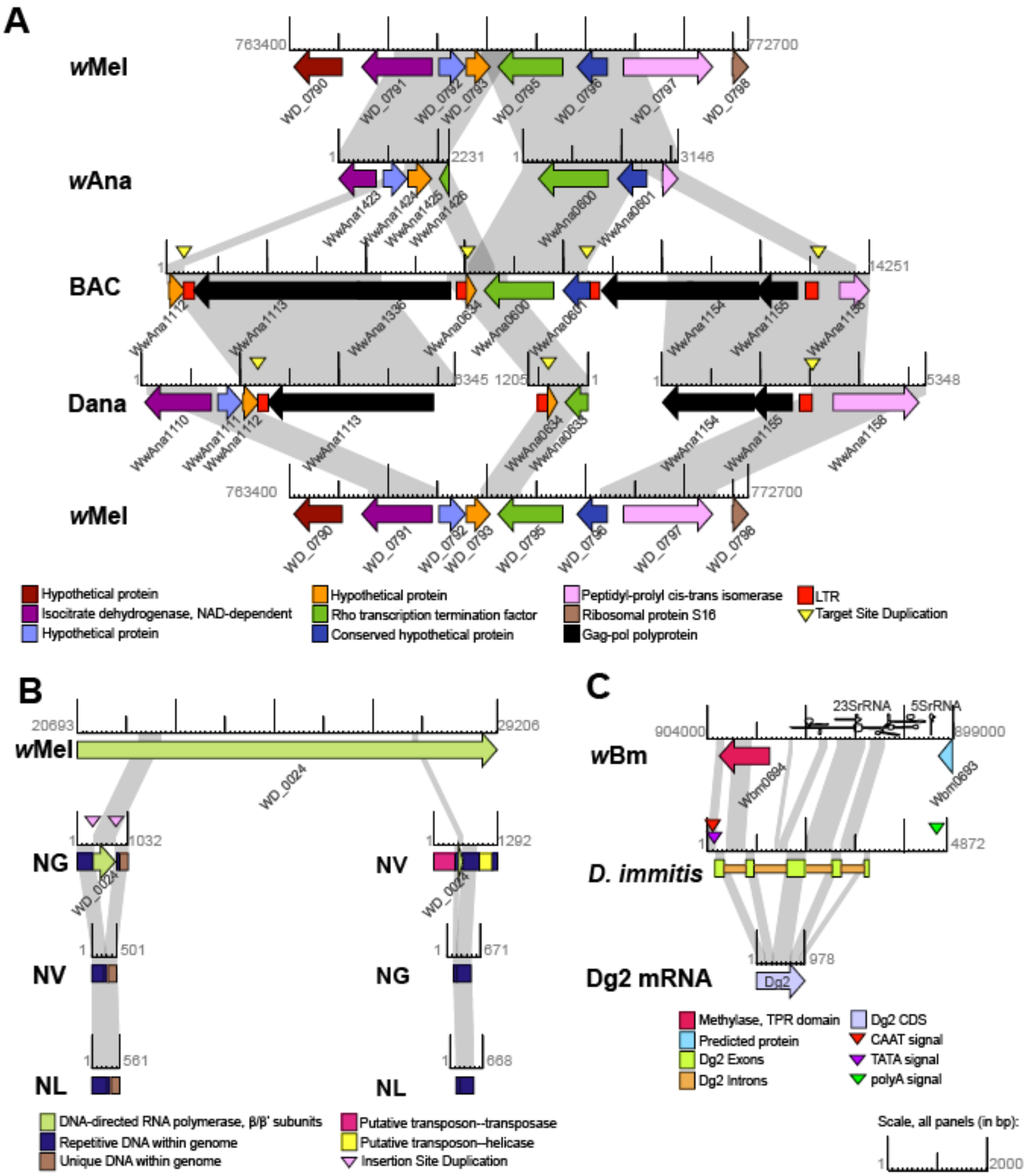


Figure A-2. Schematics of *Wolbachia* inserts in host chromosomes. A. Contigs containing *Wolbachia* sequences generated from the *D. ananassae* Hawaii shotgun sequencing project are segregated into sequences coming from the endosymbiont (*w*Ana) or from the *D. ananassae* chromosome (Dana) based on the presence/absence of eukaryotic genes in the contigs. These are compared to those from the reference *D. melanogaster Wolbachia* genome (*w*Mel) and a *D. ananassae* BAC. B. Fragments of the *Wolbachia* gene WD_0024 gene have inserted into different positions in the *N. giraulti* (NG) and *N. vitripennis* (NV) genomes with unique insertions in each lineage, including *N. longicornis* (NL). C. A region in the *D. immitis* genome that is transcribed has introns similar to *Wolbachia* sequences. All matches in panels A and B have >90% nucleotide identity; those in panel C are >75% nucleotide identity.


A
wMel
763400
772700
WD_0790
WD_0791
WD_0792
WD_0793
WD_0795
WD_0796
WD_0797
WD_0798
wAna
1
2231
1
3146
WwAna1423
WwAna1424
WwAna1425
WwAna1426
WwAna0600
WwAna0601
BAC
1
14251
WwAna1112
WwAna1113
WwAna1336
WwAna0634
WwAna0600
WwAna0601
WwAna1154
WwAna1155
WwAna1156
Dana
1
3345
1205
1
1
5348
WwAna1110
WwAna1111
WwAna1112
WwAna1113
WwAna0634
WwAna0633
WwAna1154
WwAna1155
WwAna1156
wMel
763400
772700
WD_0790
WD_0791
WD_0792
WD_0793
WD_0795
WD_0796
WD_0797
WD_0798
Hypothetical protein
Isocitrate dehydrogenase, NAD-dependent
Hypothetical protein
Hypothetical protein
Rho transcription termination factor
Conserved hypothetical protein
Peptidyl-prolyl cis-trans isomerase
Ribosomal protein S16
Gag-pol polyprotein
LTR
Target Site Duplication
B
20693
29206
wMel
WD_0024
1
1032
NG
WD_0024
1
1292
NV
WD_0024
1
501
NV
1
671
NG
1
561
NL
1
668
NL
DNA-directed RNA polymerase, β/β' subunits
Repetitive DNA within genome
Unique DNA within genome
Putative transposon--transposase
Putative transposon--helicase
Insertion Site Duplication
C
23SrRNA
5SrRNA
904000
899000
wBm
Wbm0694
Wbm0693
1
4872
D. immitis
1
978
Dg2 mRNA
Dg2
Methylase, TPR domain
Predicted protein
Dg2 Exons
Dg2 Introns
Dg2 CDS
CAAT signal
TATA signal
polyA signal
Scale, all panels (in bp):
1
2000

Appendix B

Supporting material for the characterization of the *Wolbachia pipientis* insert of *Drosophila ananassae* (Hawai'i).

Table B-1. Results of hybridization experiments with *Wolbachia*-specific genes on the *Drosophila ananassae* (Hawai'i) BAC library. The BAC library was constructed with a non-cured strain of *D. ananassae*, thus BACs may contain endosymbiont DNA or *Wolbachia*-inserted DNA. Using Finger Printed Contig (FPC) analyses and BAC-end sequence data, we chose seven candidate BACs for full sequencing.

Marker	*Gene name*	*No. Positive hits*	*No. Clones in FPC*	*Contig(s)*	*Candidate BAC* [d]
WD0024	rpoBC	89	81	8,16[b]	-
WD0132	pheS	42[a]	42	undetermined[c]	-
WD0146	gatB	70	19	1[c]	93L12
WD0229	sodium glutamate symporter fam, putative	24	21	2[b]	17L08, 89K16
WD0359	UvrD/Rep/Adda fam.	73	32	2[b]	-
WD0382	hypothetical	62	27	2[b]	-
WD0413	aspS	38	33	2[b]	-
WD0461	pyrF	47	47	2[b]	18H02
WD0496	rare lipoprotein A, putative	60	57	2[b]	51E18
WD0500	typeII secretion system prot, putative	42[a]	0	-	-
WD0549	secA	42[a]	42	undetermined[c]	-
WD0723	ftsZ	50	9	4[c]	-
WD0793	hypothetical	42[a]	42	undetermined[c]	-
WD0797	peptidyl-prolyl cis-trans isomerse D, putative	42[a]	42	undetermined[c]	-
WD1005	hypothetical	39	39	1[b]	15C04
WD1063	wsp	17	17	2[c]	-
WD1148	phage integrase family site specific recombinase	110	97	1,4,5,7,6,22,27[b]	-
WD1173	type IV secretion system protein VirB4, putative	42[a]	42	undetermined[c]	-
WD1238	fbpA	30	13	3[c]	41H15

[a] Clones were retrieved using pooled probes from *Wolbachia*-specific genes. Real No. of positive hits is undetermined; 42 of the pooled hits were used for FPC.

[b] Contig numbers correspond to FPC analyses performed in June'08

[c] Clones were retrieved using pooled probes from *Wolbachia*-specific genes. Positive hits were not fingerprinted.

[d] Library name is DA__Ba. Available from Arizona Genomics Institute.
http://www.genome.arizona.edu/

Appendix C

Supporting material for CHAPTER FOUR.

Supplementary data C-1. Genomic DNA (gDNA), coding nucleotide and deduced amino acid sequences for the *Antennapedia* (antp) gene in *Apis mellifera* (am), *Bombus terrestris* (Bt) and *Nasonia vitripennis* (Nv).

Gap data

A 281 bp gap was closed in SCAFFOLD23 of Nvit 1.0. First base pair in gap would become coordinate 1468011. Last base pair in gap would be coordinate 1468291. Information submitted to the Nasonia Genome Sequencing Consortium through the Nasonia annotation tool hosted by the Human Genome Sequencing Center (HGSC) at the Baylor College of Medicine (BCM). (Stephen Richards, Unpublished, HGSC, BCM).

```
>antp_nv_gap_281_bp
TTTTTCTCTTTGGCGACGACGACGACGACGACGACGTCTGCCTCGAATTTTTCTCTCTCGCTCTCTTCCCG
GGACGAGCCGCGTTACGCTCAGGCGAGCGCGTGTACATATAGGTATATATACGCGCTATCTCTTATTTTCT
CTCTACTCGAGCCCCCCTCGCGCGCTCCGTCTCTCTCTCTCATCTTTTTTATACGGTTATCCCTCGTTTCT
CGTTTCTCCTTTTTCCCAGAGAACGTCCGGTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
```

Genomic Data:

```
>antp_nv_gDNA_FGENESH+_+_Apollo_+_manual_annotation_12086_bp
ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGATCTGCGCAATGGCGGGGTGGAGCATCCGCATCAGCA
CCAGCAGCACTACGGCGCGGCGGTCCAGGTGCCCCAGCAGCAGCAGGCCGTGCAGCAGCAGCCCCAGCAGG
CGAGCGACCCCTGCGACCCGTCGATGCTGCGCCAAGGCGTGCCGGGCCACCACGGCTACGGGGCCGCGACG
GGCCAGCAGCCGGGCATGCCCTACCCCCGCTTCCCGCCCTACGACCGCATGGACATCAGGAACGCGGCCTA
CTACCAGCAGCAGCAGCAGGAGCACGGCATGGACATGGCCAGCTACCGGGCGAGCTCGCCGAGCGCGGGCA
TGGCCGGCCTCCACATGGGCCACACGCCGACCCCGGTCAACGGCCACCCCGCCAGCACGCCCATCGTCTAC
GCGAGCTGCAAGCTCCAAGCGGCGGCGGTCGACCACCAGGGCAGCGTCCTCGACGGGCCCGATAGTCCGCC
GCTCGTCGACGCCCAGATGCACCACCAGATGCACCCCCAGCACACGCACATGCAGGCCCAGCAGTCGCACC
CCCAGCAGCAGCCCCAGCCTCAAGCGCCTCACCAGCAGGCCCCACATGCAACCCCAGCAGACGCAGCAGCA
GCACATGATGTACCAGCAGCAGACGCAGCCCCAGCAGCCCCAGCCCGCGGCGATGCACCCCCAGCAGCAGG
CCCAGCAGCAGCAGCACCAGGGCGTCGTCGCCTCGCCGCTCGGCCAGCAGCAGCCCGGCACGCCCCAGAGC
GCCGCGCCGACGAACCTGCCCAGTCCCCTCTACCCCTGGATGAGGAGTCAGTTTGGTGAGTAGTGCAGTCT
ATACTATATAATATCGCGCATCCGAGTGCTCGGAAATAGGCGCGATGTCTCCAGCTTTCTACGTAGGCGCG
CACGTACGGCAGCCGGCGGCGCGTGCTCGAGTTATTTATTGCCCCGACCTTTTAGCCCGACTCCGCGCCGC
GGCTGCCGTGCGTTTCACCACCGCCGCGAACCAGTTTTGCGAGCGCGTTAAAGTTGCGTCGACCACCGGTG
CCGAAACTATGGTTACGCTCTCTTTGCACGATTTTCACCTCAACCCCGCATGCGAACGAACTGGATTTTTC
GAAAATTGCGCAACGCGCCTTCGAGCAGCGCCCGAATTCCAGCGCGAAGGGCTGCTCGGCTTCAAGGCCGC
```

```
TTTAATGTATCCCGGTCGGAAATCCCGGCGCGCTGGCTGCGTACCGCATACGAACCGGTGAATGATCCCTC
GTCGGCACAATGTATGCAACAGCGGCGAGCCTCGACGCAACGAGCTCCGGGATTTATGGCAGCCATAACGA
AGTGATAATAGGCCGGGCTCGTAATTTATACACGCGCAGCTATAGCCAAGCGGCTGCAAGTATCGGCAGGA
GAGCCGGTCAAAGAGGTGCACAGGTTGAGGGACCCCCGCTGACTTTGCTCTCATTCCGTCGCTCTCCCCGC
TTGCGTCTAAGCCGATGACTCTATGCGCTCGTAACCCGGACGTAATTTACGCTATGAGTGACTCGACGGTT
ATCCAGAGATTTTCGGGTATAGGTAGTGTCTTGCGACAGCGACGGGGGCTCTCAGTCGCAGTTCGATTTCC
GCCGGGGTGAGATATGCGGACGCGAAAGTTTCCGAGTTGCCGATCGTCGTGCGAGTGTTTGCTGCGCGTTC
GGTTTCGATCGAAAAGTTTATTTAACTCGGTGCCGAATATTGTGATTAAGTGAGTGCGTTGGTGCTACGAA
AAAATTTCTCCGATCACGAGCACCTCGAGGATGAGAAAAATCGGTACGGCATTTCAAGCTCAAACGCGATC
CTACGCCACTGCAATTTCCAAATTCGAATCTCCTCAAAGGGAAAACTTCGCTCGTCCCATCTCTCGGCTGC
TCTCGGAAGCTGCGGGCCCCGCTAAAAGTGCCCTCCCGGTACACTCGCCGATCGTCGATCCGTATAGGGCC
TACATACCGGCAAGCGAGAACGTCTGCACGCATTCCCTGCATAACGACTCTCCCCACACACGGCCCCCCAG
TATAAAACCCGCCTCGTCTGCACACACGCCGGTTAGTTCTCACATATACCAACACACCGGCCTTGCCAGTC
GAACACCCGTCGAACTCGAGCGGGTTCCTGCATCGGCTCACAGATGTCGGCACCTGGACTTTTAGGACGAA
AAACCTGCTTCGATCCGTGTGCGGCCTCACCTTAATTTTGACTGCAGCGTGTATTTTTCGTACATATATAC
ATACCCGGGTAAATTCCTCGGGCAACGCGATAATCGCGAGAATCCCGCGGCGAGACTCTTATTGACCAAGC
ATCCCCCATGAATTATGCAAAGCCCAATTTCCTCCACGTATCGATGCGCGAGCGCGTGTGTTTATCCGATA
TACGTGTCGGGGCGAGGTTTTTGGAATCGACCGGTCGTTGCGATGGGTGGCGGTGTTTTTTCTTCTGTTTT
CCTGTTTTTATACTTCGCCACGCTATGCTTATCCGAGGTAAATATGTAAACGCGTGGGATAATGCGTGACT
GTGTGTGTGTGTGTATATACGTAGCGAAGCGACGGCTTTAATAATTAGCGGATGAACTTCCGAAATGGCTG
TTTCACGTGATCGAGGAAGGATTGAGCTTTCTATTTCGGCTCTGTACTGCTGATTCGATTAGGAAATTATT
CATGTCATCGAAAAACTCCTGAACACACACTTCTTCGAAAGCAGGAAGCCGCGGCGTTAAAAATCAATCAT
TTCCCGTGCGCGGCAACGCCCGGCGCGAAACAGCCGCGCACAGAGATATCGCACGTGCCTCCTGTATCCAT
AACAAATCGTTACTTAACGTGTATTAAAAATCAGATATCCGCTCCGCCTTGTGCTATACACACAACGAATC
GACTTTTCCATTAAAATTCACGCCGCTCTCCTTTTTGCGGCCATTTACCTCGACGAGCGAGAAAGGGAAGA
AAAGCCTAAACGCGCATTCGAATCGAATCTCCGCCGGCAAAACGAGAAAATAATTTCGCGCCTTATCGCGA
CGCGATATCGGTTCTCGGGGCTGTAAAGCCGGGGCATTATACAGGCCGGCACGCGAGTCGCCCATTACTCG
CGAATACACCCCGGCTATAGATTAATACTTGAAATATACGACGAGCTCGCGGTGTAATTTAACGCGACATG
CGCGCACCTATTCCAACTCTATACGCACATGCCGCATGGGGTATAGTGTATAATAATATGCATTAACGACT
CTGAGCGCGATGTACAAAAGCGACGACTGTCATTGGGCGCTGACGGACGCTCGAAGAACTGCCCCCCTCGA
TGATGGACGTCGCCGCGACTGTCATTATGTCGTACTGTACCGCGTTCTCCGTATAATATACCGTAAATAAA
TCAACAAACGTCGATTTTTTCACATCGAACAGAACTCGATCTCGCGCGTTTAATTTAAATTCGAAAACCCG
CGTGGCATTAAAAAATCCGATATCACAGCGAACTGTAAAGGAACCGCGTCGGCGTCTTCTCGCACACACAC
ACACATGCACACATACATACATTTAACACACAGAGCGAGAGCGAGAACCTGCAGCAACAACGGTGCTTTTA
ATATGCGCGAGATAATTGCGATGTGCTTCGTCGGGAGATTTATGGCATCGCCGCCGCTCTCATAAATATAA
GTGCGCGCGAGCTTGGCTCTCTCGAGAAAGCGATATACGATCGGAGCCGCGCGTCGCGGCTACCGTTAACG
CGAGCCGCACGGCCGATTTCGCGGATTTCACGGATCCTTCGTATTACACGCGCGCTTTTCATATACATCGT
GTTTCTTTCATTTTTTTTATCGTTCGATGCGTTACTTTTCGCGCGCGGATTAATTTTATCGCCGCGCACGC
TTATTATTATCATTATTATTACATACGAATAATGAGCTTTCGCGCGCCGTTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTCTCTCTCGATCGGATATTGAGATTGATATAATTTTTGGTGAAATTTTTCGGCGCGCGCGCG
TTTATTAAAAGCGGATTTTTGAAATTTTGGCCTTCGCAGCTCGCTTCGAGAGATTAATCATTATTATTATG
TCTGACGCGCTGCTCTCTCGCGTCGTATCGAGAATTATGGTTTGCATTTTTCACGAAGATTGCGTATGCTA
AACTTTTGTCAACAGTCGTTATTGGGGTTATTCATTTTTGTCGAATTTTTCATATAGAGAGCAAGCAGGAT
GATTCAAAGTCGTCCCTTTACAATATTACAGGCCGAAACTTCTCCCAAAAACACAGCGTTCTTCCATTTTC
CTCGCATACGCGCGCCGCATACGCATAGCCATCACACGTAAAACCACTCGCGGCGCTGCACCATCTGTCCC
TCGGCCAGCAGAGCGGGGCTGCGCACAAAGAAGGAAAAGCGCGGTCGGAACTTCTCCCCCGAAAACTTTCT
AAAAATTTTCCCTCGATATTCCATCCCATAATTACAACATCCGCGCGCTCGCAGCCGCTGCTCTATGCTCT
CTCTTTCTCTTTCTCCTCTAATTATTTCCTCGTGATTCGGCTGCTGCAGCAGCAAGAGAAAGGATATGGCT
GCCGCTGCATCTCTCTTGAGAAAGGGATGACGGGGAGGAGCGAGCGAGAGAGAGAGAGAGAGAGAGAGAGA
GAGAGAGAGAGAGAGAGAGAGAGAGAGAAAATGAGCCGACCTGTGCACGGAAGAGGGCATCGCATTAGAGAGA
AAGAGGGAAGAACAAGACGAGGAGCGGCAGACGTAGAAAACGATACACACGCGCGAGGGAGGAGAGAAAGG
AGGGGGAGGGACAAATGACGATGATGGATGAGCCGCGCACGAAAAGCTGCCCCGGGAACGTGTGTGCGCAC
TTCGCTTTTTGAGAGCCGCCATGTCTGCAGCGAGATGCTTTCTTTTTGTTTTTCGACTTTTAATCTTTAAC
GCTTTTTGAGAGCCGCTTTATCTCTCTGTCCCTTGAGTTCCTGTTTTTTTTCGCCGCCTGTGTTTCTCCAG
```

```
AAGGAAGCAGCCGTTAATAGCACGTACGCGCCGAATATCTAATTGTGTTCTTTCGGTTCGAAAATAAGCTG
ACGCCGGAAAGTTTTCGCTGAATCTCGAAAAGTTGAAGAGAAACAAAAAAACTCTCCGATATCCCATAATT
CAAAAATTCTCTCGAATTCCAGTAAAACTGATTCGTGGTAAAAATCTCCGGTATTCCGAGGGGCTTTTTCC
TATTTTCTCTCTTTCTCTCGGCCTGCCTGCCTGCCTCCAGTCATTCTTCCTCTCCGGCTTTCAGGCAACGC
GGCGACTCGCCGCCGCACGCGATATGGCCAGCGAGATATGAGTGCAGCGCTCGCGCGCGCGTGCAGTGAAA
GCTTCTTTTTCTCTTTGGCGACGACGACGACGACGACGACGTCTGCCTCGAATTTTTCTCTCTCGCTCTCT
TCCCGGGACGAGCCGCGTTACGCTCAGGCGAGCGCGTGTACATATAGGTATATATACGCGCTATCTCTTAT
TTTCTCTCTACTCGAGCCCCCCTCGCGCGCTCCGTCTCTCTCTCTCATCTTTTTTATACGGTTATCCCTCG
TTTCTCGTTTCTCCTTTTTCCCAGAGAACGTCCGGTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTCTCTCTCTCTCTCTCTTCTCGTCTCCCGCCGCCTCTCGAACAAAAAAGAAACTTTCCTCCT
CTATTTGATTCGAAAGCCTGAGCTGCGCGCGCGCGAAAGACAGTGACTGCTTTGTTCGCAAATTTAATTAA
GGGGCGCGCAAAGCTCTCTGCTGCTGCTGCTGCCGTTGCTGCAGCGAGAAGAATTTTCAGCCCCGTTTTTA
CACGGCCTTCTCTTCTTTGTTCAGCCCTCAGCTCTCGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTTGTAGAATCTTTTTGATGAATGCGCGCCCCGGCCCTTAGAGCATAAATAATTCAGAAGATA
TTAAAATAATGCACGGCTGAGATTTCGACACACACGCAGAGATAGAGAGAAAAAAACACAACGTGTTTACG
TATCCGAATAGCGGCGTTTAATGATAATACGAGAGGGCTATTTCTCGTATACGCGCGAGCGTTTATAATAA
AGCGATGTTTTGCGCGTGCGACATAAAGTATCAAAGGCTGGGCAGAAATTAAAGGAATCCCTTTGCAGTCG
CGCTCGCTGGGACGATGATGGATCTGGTGAGCAGCCCGCGCGCTCACTCTCGCGAGCGAGCGAGAGAGGAG
GAAGCAGACGTGAAAGAGGGGCACACGTGTCGCGTCTAAAAATAATTCGCCGCGCGGCGGCCCAAGAGAGC
CGCACGTGAAAGGAAGTGCGCGGAGAAAGCTTTTCTCTTCGCTCTCTTTCTTCTTTTCCCGATGTAGGAGG
AAACTCGCCCATCGGAAGGCAGGACCGGGCTAGGGTCGTGTGCACGGAGCAGTCGTGCGCCGTGAAAAATG
AGAGATCGATCCGCCTCTGAACTTGCGCGTGCAAGCTCGCGCGGCTGCCATCGCGTGTTTTTATCTGGCTT
TAACCGCGAGAGCTTGCAAAGAGGGATGCGTAACACACGATAAATTAATAGAAGCTATGCAGTGGTACACG
CGTAGGAGAGGTTCGCCGGGCCTAATTGGACCAAGCCGCGCCTTATTAGCGGTTAACATCCGGCTAGCTGT
TAATCTGAAAAGTATATCCTGCGTATCCTGTGGCAGCTATCCGATGCTACTCGCCTCCGAGATCGGCAGTG
ACTTTCATAGACAGAGGGAGAGAGAGAGAGAAAAGTCGCGGCCTATAATAAGCCAGGAGAGTGTGGGCAGC
CGATAGGTATACGTGTGTGTTTTGTTCCACCGCGCTTGCCTTGCGCTCCGTTCCTTCGGGTGAGAGAGAGA
GAGAGAAAGAGAGAGAGAGAGAATAAGGTAGACCCATAAATCTCTCGACATACTCCAGGGGACATTTGATC
CTTGCGTAACCCAACGCCGTTAATTTATTGCCCCGCGTTTAAAGTCACCTTTGTGCAGTGTATACCGTCCG
CGAGAAACCAAGTGCGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
GAGAGTACGTAGAGGAGAGGAGGCAAAGACAAGGAGAGGCGAGAAAGAGCGCCGATCTCAAACTCCACTCA
TTTACTCGGATAAACAAGTTTATCTCAAAGCGGGTGAGCTTTTCTCCGAAGCGTCGTCACTACAAAGCTGC
AAGAAACAATATGGCGCGTATAATTACCGGCTCTAAAATCGCACGCTCCGAATCTCCACAATTCGCATCGC
TCGCACTCGAGAGAACGACACCCGAGAATCTCAGTCTCTCCGCACTGTGTGACCCAGAAGCATTTCTAAAA
TCCGACGATCCGCCGTGGGCCGCAATAAATTTCAGTCAAACCGCGCATTCTTTCGCGCGTCGTTAATCTTT
ATCGCACGCGCCGAGCGCTTTTATTGTTAACGAGATAGCTGCTTCCTCGCGTATATTATAAAATGCGTCAT
CGCTTGTTGGCATACGTCTCTCTCTTGTTTGCCGCGGCCGAGACAATGTTTCCGCTGCTGCACTTGACTCG
TCATTAAGATAACGCGACTCTTGAGACGAGCACACTCGAGCGGCTGAATTAATCGCCGTGTTTTGTGGACG
ACGCGAGTGCCGTATGTCAATAGCACGTGTGAGGAAAATCGATCCTGGAATATATTGTCTCTCAATGAGAT
TGACGTCAAACACTCATAAATAATGCTGGCTTCTCACAATGCATACAGAGAGCAAAACAGTCGCGAGCGTC
GCGAAGATTGATGCCACGAGTCAGTCGGCGTTACGAAAATATATTCCTGGAAAAGAGGAGCGAAGAGTCGC
GCAGCGCCACACTGCATCACGCTCTCGCGCAACCTCTTTAATTAACCCCGCGTGTGCAGCTCTCGGCTCTA
ATCTCGGAGCGAGCGCAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
AGAGAGAGAGAGAGGTAATAATAATACGCGCGAGACTGTCCGATCCGTCCGCGCACGTCTCTCTGATGGGA
TTTTAGACTCTGCTATAGCAGTCGGTTTTAGGAGATTCCGCGCGGAGAACGATTTACGAGCGCAACCTCTC
GGTTTGCTCTCAGCTCGGATGGTTTATCGCAGTTATCGACTATTTGCTTTTCTAATGCCGGCGGTTTGTAA
ACACTCGTGCGCAAAAAAATTATTGCAGCTGTAGAGTCGTCGCACTTTTTCTTTGCGACTTTTTAAGCGTC
TCTTGATTTACGCCAGTCGTTCAGAGCATCTTCCGGCACGTGTGCGAATCGAATCTCGTCAATGTACTTTC
ATTTGGGATAAGGTAATTAAACGCGATTGACACGGCTCCCTCGATAGGATTTATTAGCTCGGATCCTTTCT
TAAGGTACCCGCCTTTCTCGCGCCCATATTTATGGAATTTCTAATGCCTCTAAATCATTTCGACTCCTGAC
GTGGTCGTACGCGTACAATGTATATCGGCTCTACTGTAAATATAGTCTGAATAAACATTTCCAAGCTTACG
CTTCGACGAATGCGCGCATTATAAAAAGCGGCTTATAATCTCGTAAAAACGTGGGGTATAGAGCTTTCAAA
GAGCGCATGCATAATTAATTCTAATTGCCGCCACAGCCAAGACACTTTTCTGCTCGAGAGCTGCAGTGTCT
TGACATTCGGAATGTCCGTGTCCCGCTTGATGAATACGTTACCTCGCACCGTACAGATCTTGGTCAATAGC
```

```
GACGCTGGATCAAGTTTTCCTCGTGTATGCTCAGAGGAAAATCGCATTTTGATTTTACCGAGATTCTTCGA
ATTCCGAGAATCTCATCGTGGTATACCAATTAGCGTGCTGTATTAGCTTTCCTCTTCCCAAGTTTTCTGGG
GCGCGGATAGGCGCTAGAGATAGAGCTCGAACCGAATCAGGGATTCAAGCATATAAGAAAAGAGCTCTCTC
TCTCTCTCTCTCACACATTTCTCGGTATTTTTTGTCGCCGCTCTCGACACTCCGGCCTTGTAACCGAAGAA
GAAGCGCCTAGAGAGATTGTTAGCCAAGGAGGGAGCCTATAGGGAGATCGGTATGTTGTTAATTTGAGAGA
GATAGAGAAGGAGTAGAGACGACGGCCCGAGGGAAGGCTCATACACACACACACACACACACACAGAACGC
GTTCCCAGAGAGCCTTTAGCCCAGTTCGCGCTTGCGCCAGGCTATCTTTTTCCCTCTTTCTCCGCTCGCCC
TTTCTCCTTCTTCATTTTTTTTTTTCTCCTCTCGGCCGCCGCCGCCGGGAGAAAGAGAGCGATATTCTGAA
ACCCCGAAGAATCTTCTCCGTCTCGCTCGCTCTTGCGCGTGTAGAGGGAGAGAACCGCCATCCGTCTTCAT
CGGCCGACAGCGCGACGGGCGGCGTTTGCCTCGAATGCTCACCAAGGAGTTTTCTTCTTCTTCTTCTTTCT
TGCAGCTTTTTTCTTCCGCCGACGTCCTTCTTCTCGCCGGCTCATCCCACTCTTCGAATCTCCATGGCGCT
GTGGGCGCGCGTATCCAACCCTTTTTTTCGTATTTATATACGTACACGGGCAGGCTTCGCTCCGGCGAGAG
TATTATATTATATGCGTGTCATGCGCTGGAAGTTTGTTTGCGTTACAGACGCGTTGCGCAAATACTGTAAT
TACTCGATTGCGCTGCGTAATTATTGTTCCGAGAATGCGTGTAAGGTCTGCTTTGGTTTGTACGAAGCTCG
TGGATTGAAAAACCATCGCGTAATAAGCACGTGTCGACGGACAATGGCTGCGATGGGTGTGATTAAGCCGC
GGTGTGATCATACGGAGTTTTGGGTAACGGAAGCTTTACTTAAAATCGCATTACACTCGAAAGAATGTTTT
TGATTTTGTAATTTTGTTTTTTTTTATCATCGAATAAATTATCTCTCAGCTTCTCTTCCCAGCAATCGCAA
AACGAAATTCGCGAAATCAACATTTTAAAAACTCGTCTTACGCTTCTTCTTCTTCTTCTTCTTCTTCGGCT
TTTTCTGCGCATACTCGAAGATTAAACTCGCAGCGGCTCCCCTCTCGGCGCAATTAATATTCCGATTCGAA
TATAAGACTCTCTCGGAGAATGCGGTCCGATCGCCGAGGCGATAATACGTCATACGCGAGCTCATACGTCC
TCGTCGCATTACACGGCTAAATCTCCGAAAAATGGCATTATTCAGTCGGTGCGCGGCGATAAGGATTATAT
ACGCAAAGGCGAGGACGACTCATCGCAAATAAAAGCTCGCGAGGAACAGCTTCTATGCAAATACGCCACGG
GGATAAGGAGAATTTCGTTTTTGCAGGTCGCTTTTTTCGTCGAACCGTAAGCGAACACGAGTAACTCCCTG
CAGTCTTGGAAACGTCCGCTGTAGTTATATTGATGAGCTGCGATACATTATCGCGGGAATCGCGCTGTTTT
GCGATTCGCTCGTAAACTAACACGAGTTAGGATATAAACATTTAACGAAAATTCGCGATTTAGACGACTAT
AATTCACGAAGTTTGATTTTGTATACGTAAACACGGCTTTGAGCGCAGTATCCTATTTGCAAAACAAAGCG
ATATTAATGTAAGACAAGCCGCTAAAAATACTTTCCTCGAAAACCGTAACGTCTCGGAAATTTCAACCGTC
AAGATTGAGTTGGAGCGAGGAAAAAACGTCGAGAGATAGGAGAGAGAGTGAAAAAAGCATTTCTCGGACTT
TTTCTCGGAATATTCGTGAGAGTACGAGATTCGAAAGAGAGCATTATTAATGGGACCTCGGAGGAAGTGAC
GGCCCAGCGAGGCTCGAAAACACCGAAGAGCGAGAGAAAGAGACGTCGCGCGCATACAGTCGTTGCGTTGG
ATTCCCCGACTCGTGTCCGCAGCAGCGACTATGATATATAAAGTCCTTGTGGGTCGAATGACGGTAAACAG
CGGTAAATAAACGTCCGCGATAGAGGATGACGCCCTCTTACCTGACACTTTCTGAAATCCCACGCTCTTCG
AGTCTCGGGAATGTTGCTTCTTTCGAGCGATTGTTTAAAGTATAGTCGAATTCCCTAAATCGGGCGAGCGA
AACGCCAGGCTGCTTATAGTGTCCATAGTGATTTGCGCGCGCGGCTGGATGCGAGAGGGTTATTTCTGCCT
CGAGTCCAAGAATAATTGTCAATTACGAGGCTCTGATTGAAATGACTTGCAAATTGAATAACCTGTGTTTG
CTCTGCGCGGCGTCGGCTGATAAATCGCCTGGCGGTGGCATATTGGTGTTATTGCCAGTTTCTCATACACA
TAACGTATGTACGTGTGTACGTTTAGACCCACGCGCGCAGGCTCGGATAATTTTAACCGCGTCAGCCGATA
TCGAGTATTATTATACGTCTTATTAAAAAGCGAGAGTGTGCGTGCCATAATAAGCTTATAAATTATTTTTA
CCGTCTCTCAACCAGCGCGGATCGTAAATCTTGTCGCGAATCGCTAGAGATAATCAATGATCGCTCGCCAA
GTAGCAGCATCGTTACTTTGATCAGCTCTTATAGCTATATCTCATAGGAGTTTAAATTTACGAGCGATTCA
ACCATCAATCCGCGTTTGCTCTGTAAACGAAATCTCCTCACTCAATTATACATCGAACGATAATATCATTA
TACGATGATGACTAACGAGAGAACCTTCTTTTCGCTTGTTTTCCAGAGAGGAAGCGTGGCCGGCAGACGTA
CACGCGATACCAGACCCTCGAGCTCGAGAAGGAGTTCCACTTCAACCGCTACCTGACCAGGCGACGCAGGA
TCGAGATCGCGCATGCGCTCTGCCTGACCGAGCGCCAGATCAAGATCTGGTTCCAGAACAGGCGCATGAAG
TGGAAAAAGGAGACCAAGACGAAGGGCGAGCCGAACTCGGGAGACGGTGACACCGACATCTCGCCCCAGAC
CTCGCCCCAGGGTTGA

>antp_bt_gDNA_FGENESH+_+_manual_annotation_11883 residues.
ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGACCTGCGTAATGGCGGGGTGGAACACCCGCATCAGCA
TCAGCAGCACTACGGTGCCGCCGTCCAGGTGCCCCAGCAAACGCAGTCGGTACAGCAGCCATCTCAGCAAA
CCGGGGATCCATGCGATCCTAGTCTCCTACGTCAGGGAGTGCCTGGCCATCACTATGGGGCCGCTGGTAGC
CAGCAAGATATGCCTTATCCGAGGTTTCCTCCGTACAATCGGATGGACATGCGGAACGCGACCTATTATCA
GCATCAACAGGAGCACGGCAGCATGGACGGGTTGGGTGGTTACAGGTCGACGTCCCCGAGCCCCGGTATGG
GCCACATGGGACACACACCGACCCCGAACGGACATCCGTCCACTCCTATTGTCTATGCGAGTTGCAAGCTT
```

```
CAAGCGGCCGCGGTCGATCATCAGGGTAGCGTACTCGATGGACCGGACAGCCCGCCATTGGTCGAGTCGCA
GATGCACCACCAAATGCATTCGCAACATCCTCACATGCAGCCGCAACAGTCACAACATCAACAGCAGCAGC
AGCAGCATCAACATCTTCAGGCGCAGCAGCAGCACATGATGTACCAACAGCAACAACAAACACAGGCGGCG
TCGCAACAATCTCAGCCTGGCATGCATCCGCAACAACAACAGCAACCTCAGCAACACCAAGGGGTGGTCAC
GTCGCCGCTTAGTCAGCAACAGCAGGCCGCTCCTCAAGGTGCGGCCACTGCCAACCTACCAAGTCCGCTCT
ACCCGTGGATGAGAAGTCAATTCGGTAAGATGGACTTTCAATGTAGATGCTCAAAATGTCTGTTGCGTGCG
TTAAAACCGTGCAAAAGTATATACAAATTCATTGATGTCTCTGATGGATATGATATTACTCATATTACAAT
AATACGCGATGTACTATATCGAGAATTTGAAGTCTCTTGAAATACACAAAATAGTTTTAGCTGAATGAGAA
TAATAATCGAATTTAGTGAAAGAAGTTTTCACGGTAAATACTGTCCAGTAGAATCATTCGAGAACGTTACC
TGGTCTATAAATTTAAAGGATTAGGAAGGCCAGTAGAGTAGTAAAAATCGAGAAGCAACCATGTTGAATCT
TGCATTTGCATGCGAATTACCCTGCTTGTTACCTCTATTTCTAAAATTGTACAATGAACGTTACTTCGAAG
CAAAGAATCTTCCTATCTCGGAGTGACAAAAAGAGAAGAAAAAGAAAGAAATAAGAGAAATGTTACATTCG
TATAGTCGGAGAGGAGATAATACGTGAGGCGGTTGTCGAAAATAGGCTCAAGGATTTTCCCGCGTGACCAA
AAGGTTCGGGTAGAAAGCAAGATTTCTAGAGAGGACGGCAGGTAGTCGGCCGCTTTCAGATTTATGCAACA
GTAGGTAAAACTAGACCACATTCGTACGGTTCGCAAGCCAAGATTGGCATAATGTCTCCAGTTTCTAGGTA
TACCGACATCCCGAGTTATTTATTGCCCTGACCTTTTAGCTCTCGCTGGTTTCACCACACGTAAGCGTATA
CCTCCCCATAAAAGTTCTCTTGCTGACTGTGCCCGTGTTGGCTCCACGGCTTGGCAGAGTCTCCGCTCATC
CACTTTTATTCCCATGATAACTCGCGGCTGCTTGCTTCCACTCAATGGTATGCTGAACTAGTTCGCTTGTT
TTCACTCACCTTCCCTCGTCTGTCTAGTTGGACAGGAATCTAGTATATTGAACGGCTACTGCTTCTTCTTT
CTAAATAAATGCTTTCTGCGAAGACGAGGGCAATTTCTATTTAGAATTCGTTCTATTTGACGAAATTCCAG
CCTCTTCGTGATCTTTATGAAGCTGAGCGTAATATATTCAAGGGAACAAGAAATACCACCTGCTTTTCAAA
TATAGGGTTGTAGCGGAAGAAAAGATTTAGGAAATCTGCTGGAGTCGACCGTGACTGCAAAGAATCGCACG
GGTAATTTCGGAATAATGAGTTGGGAATTTATGGCAGCAATAACCAAGTGATAATAGCCGTTTTCTTTAAC
CCAGAGGTATCTGTCCCGGGCTGGGTAAGAGCACAGGTTCCGGCGCTCCACTGTGCTTGCTCTTATTTTCT
TCACCCTTCCCCTGATTCTTCAAATCCGAGAAAATGCCGCAACCTTGCAAAGTCAGGCTCCCTCGACCCCT
TTTCTTTACTACCTTAGCACGTTCAGTTGCTCGATGAACTCTGTTCAAGTTAAATCTTCTCTTCTTCAACT
CGAATTATTACAAGTACTTACTTACATATATCGACTTTCAAGATGTGAATTTTTTTGCTGCTCTTATAAAA
TCGTTCACAAGCTGCCACCTGTTCTTTACTGGATATTTTTCACGTAAGTGTAATTCGAATTTTTATACGCT
TTTTACTGTTTATTTTTCACATTGGATATGGAGAATGTAATAATCTTACGAAACCATCTCTCGGTTCAATT
CGAAAGCTGGTACCAACCTGCTTGATTCGCCGATAAACGGTGTGGTACACGAATGACGAGACTGGTCCATA
CTCTTCGTTGCCAGTCAAAAGAAGAAAGAGGATCGAGTGCCTGAATCTGCTCACAGATGTCGGCACGCTTA
AGTCTTTCCGATGGGTTGTCCAACAATATATGTATTCGTTCCTCGTGGTTTATTTAGTAACCGTATTGATT
TCATACTATCTTCGCTATATCTTTTCATAAGTAAACAAAGATGATTCATATGGAAAGCACGATATCCATTT
ACACCATCTATCCTCGAGCAAATCTTACAGAGAGAATCTTCTATGATCATTTAAAACCAATAAAATATGTA
GCTTTGTATCAATAATTCTCTCGGTTTAACCATTATAAACATAGTTCCATGTTGTATATGTGTATTTCGAT
TCCACACGATTCCATCCATCGTGTATTTATTAGCCATTCGAAACAAGATCAATCGGCTCTCAGATCAGAGT
CGATGTATTCTAACAAACATCCTTCATATCTGGTTTCGATACTAATTCGATCTATGATTATCCACGAGCAT
TCGTTCTTCTCTGTTTAACACTGTTCATGAGGTTTGCCACTCGATCGTATTACCCTTACAGGTCTACCAAT
CATGATAAACCTCATATCAATTTATCTAGCTATCCATTCGTGGCCGTGGTTGCCTGTTTCTTTCCACGTTG
CAAAACACCAGCTGAGCTAAGCGTGCTATAAGCAGGACTCCATGAATAATGCATGGCTCAATTTTATAGAT
GCAGGTGCTCTCCCGGGCATTTGCCCGGTGAAACCGCCGAGATTCCCTGGCACCTTACGCGGGACATTTGC
GTTACATAGGTTAGGAAATTCCGGCATAGTTGGCTGGCTCGCGAAACCCCGGCATGCGAGTGCGCATCGGT
GTCGTTTACTGCCGAACATCATGAAACGCTTCGTGAACGTGCACAACTTTCATCGTGAAAGCTTCTTCTAT
ATACAATGCCCCTTTGCTCCGCCTTTGATTTCTTACGACGACGATTTTTATTAAATATCCCTAGCCACCTG
TCGAGTCACGTGACAGCTTCTGTCGCATTACCAGTCCTAACCATTGCATAATTAGCGTTTTTGTTAAGGTA
TAGAAGGATTCGGAAATACAAATAGTTTACTTTATTTCACACGCAACGGCTATACGTATATATTACACTCT
GAAGGCATCGATAACCTACTCGCACGCACGCTTGTTGTCAACCGACTCCTCTAGTCGACCTCTGACGCTTA
CACACACATTCACATGTACAGTGTAAATGTATCATCTACCTAAGAGTTTTTATTATTCTTCGAAATGCCAA
TATGTGTTACAATCCACGATAATTCCTGTTCGTTCTGAGAAAAGGTCGACGAAAGTAGCCAAGGCAGTTAA
ATCGGAAAAGAAAAATATTTGAAAGTGAAATTCATTCAGGCAAATCATTCATTTTCACAATCAGTGTATCT
CAAGGAAATTTAACATCTTTGCTAGTAAATATTGTTTTACGTGAAACGTTATGAGCTAGTGTAAATAATGT
CCACGATTTCGTTAGCGATTAATAGCTGCGGACGCGAACCATGTTTTCGTTGGAGATACGTAATTCTGTCG
ACGTTGCAAGTTCTCTCATCCGAAGTTTGCGGTTTACGCCGACACTGAACTTCGGTAAAATTGTTCATTGT
CATTCGTATCCTGCATTCGCATGCATGCATGCACTCACATGCATAAACTTTCATTGATTAGCATTTTCTTT
```

```
TTCAAATTAATTCGCCAAGCAGATAAAATTAGAACCAATTGCGGAGAATGATAGTATTCGGGGTTCGTGAT
CAAGTTAATTAGCTGGATGGAAAAAGAAAACGCGAAAATTAGCTTGCGCAGGCTTATACGGATTCGACGAT
CGGTAATTGCGGTGTAACAACTGCAACAGCATACGCGTTATCTCAGTTTATTCCGCGAAACGCATTATCAT
ACGAAAAAATATAATATTTCTCGTAGCCAGGGCCCATGAAACACCATCGCGATACCATAAAAATAACATAC
AATGAAGTTATAATTCAGAGGTTGCAGGCACCGTGACGTTTTCACCATTCCACCGGATCTTACATTATAAG
ACAAGCAAATTTTTTATCCTTTTTCGTCATCGTTATTCTTTCGAAATCAAACTCGGAACGAAATTCTGATT
CAACGCTGATCGATTGTAATTATGTGTCGATATTGCGTCGTCGAACGTGGATCCACGTGAATTGGTCGGTT
TGATAGAAAGAAAAAGCCTATCGATATCGCTGATGCTTTTAATACGTAAGATAATTGCGATCCTTTTGCCT
GAAAAGATTTATGGTATCGAGTTCCTCCGGTTCTGCTTCGTCTCCCCTCCTCCACTTCGTCCCTTTTCTCC
CCCTCACAAATAAGCATGCAGCGGTGCAGTTATACATAGGTGTGAGGCTTCCGTTATTGCGTACACGTGTA
CGTGCTACGAGCGCCACATTTCTTATGTAATCCGAGCACTGACACGAATGCTTAGCGGACGAACAAAGTAC
CACGGCGAAGAAACCGCACTACCGTCGAATTAATCGCGCGAAATATCTACGTAAATGGCCAATAACCCGTG
CGCGTAATGAGGAATTAAAATGGCCCCTACTCCGCTCCCTGCTTCGTTTTATTTCTCCTTTTTCCTCCCAC
CTTTTTTTTTCTTTTTTTTCATACGACTTGGCTAATTGAGGGCGCAGATTTATAGACTGGACACCGTTTGT
TCATCGAGGCACGTGGCCGCGTTTATCTTTATCTTTACCAGATATTAGCGCGTATCTTTTAGTTATTTGTT
TTTTTACAGGTTCCTAACTTCTTCTCGTTTTTTCACCTTCACCTTTTTCACCGGATTTCGATACATAGCAG
CGCTGTGATTTTTTAAATTGAGTATATAGCAGCCTTAACTGCGGGCTACTCGCGATGCAGATATTTCACGT
TTATCGCGTTGTAAGATATATTCGGGTATTATTATTCCATCAAATGGTATAACTAGCCCTATTTTTGCACG
GCCAGTTTACTATCACGCAATTACACGTTACGTTTTGTTCTACAAAATAAAACGTCAGACGCTTACGGTAG
TAATAATCTACTTTTCAAGTTGAGGTTAGGAAAATCTGCAACACGCCATCTAAATGGAAGAGCGCGTGTGG
CGATAAAAATTCTTTCGCTTTTTCGTCCAAAATGAAATTTTCATTTAAAAAAAGCACAATTAGGTCGAACT
TCGATGCTTTCTTGCTCAAAGATATCCGAAAGAACGCACGAGGGTTAAGCGCAACCGCTCAACACCTGTAA
CCCTTTTCCCCCGATATGTGGGAAAGATAAAAAAAAGTCCCTAGATACCATGTGGCAGGCGTAGAGTTTTA
ATCTGCGTCTGTGCTGGCGGGGGTCGTGGCGATCAACATTCACATCTCGGAACGCGTTTCGTTCTAGTTCG
CGTAGTGTCAGACACCAGAGATTGGTGTCCGAAACGATGGCGGCTTCGTGAGATATTCAAAAAGAGATGGA
ACCCAAAGAGGCAAGAGAGGAGGAGCGTCGAGGAAATGACGTGAGGTCGAGAGACAGAAAAGGAGAAGGCG
AAGGAACGTGGTATACGGTGAAACAGAGAGGGAGCAAGTCGACCGGTGGAGGGCGTAGAAAACGTTGACGA
TGATGGATGAGGCCACAGAAATCCTATGTCCGTTTGCTTTTTGGGCCGCCGCCATAAACTCCTGAGACAAC
TATGGACGATAGCCGTTGGGATATGGGCTGCCTCACCCATCGAGCCTGCAAATGTGTGGAAATTGGGGCGA
AAAGGGTGTAAATCAATGATCTTCTTACGCGTATGAACTATAGTTCTTTTTCTACGATCGCTGGGAAGAGC
TCTTAAAATTCTGGAAATTTAATATCAAGCAGGAATTTAATATCGGTTAAGAATTCCAGATTTTTAACATG
TTTGCTTCAGTCGTTATTCCTTATCCGTAAGATAACGCGTTAAGACACTCCCAGTTTCAATATATTCGCTT
CATTTGTCACTGCTAATTCTCCAGATAACACATTCTCTGTGGTGAATTTTAAGATTCCCAGTTATCTCAAA
CCTTATCATCGTCTCTAGTTTCTCTGCGATAACTCGATGCACGAGCAGATGTAAACAAAGGGGCTGAGGGG
ACGATCGTAGATAAAGTAGAACACAACCAGCAAATCCAATAATTTTCTCTACAGATGCATACGTCGCATCG
GGCCCATTCTCGCTGCTCATTATCTCGACCTTCTGGCGCTCTCCTCGGGGCTCTCAGACTGCGTCTCCTCG
TCCGGTTCTGTCTGTTCGTCTGCGGTCGAAAGGCAAAAAAAAAGAAGAGAAAAAAAGAAGAAGGTGAAGGA
GAAGAGGGAAGGCCTGCTATCAGATACTCCGCTCTCCTCGAGGGCTTCTAAGACCCCTCTATGAGTGTCAG
TCGACTTGACGTTCTTCTTACTCCGGCGGCGACTTCTGTTGAATTCCTTTTTCTCTCTCGTCCTTTGCGAG
TGACGCTGCGCAGGTGGCTGTTCCAAGCTTACGATGCTTCCACTTGGTCCAGTTTCCTCAAACGAGAACAA
TCGCGTCCAGCTTGCAGAGGAAACTAAACCATAGTGATCGCAGTGGATAGGGCGAGCTGTCTCTGTTCTTG
AAATATTTTCCTGTTCATCGGCGATTACCAAAGACGCGTGCAACGTCCATGGCTTTGTTTTTAGATCGCAG
ATGCAGGAAGTTGGCGAAGCTTGGAACGCGACCCCGCGGATACCCGCGCCGATGCATAAACAACCGACGAA
ATTCCTGAACGAAGTTTGTCTTGACTGCGTGGCGCGACCCCTACGCTCGGGCCGTGCTGCCTGCCGGTCCT
TTTGCCGGGGAAAAAACGTCGCCTCCCCCCTTCGGGCGAAGAAAGTTCAACGAAAACAACGTGACATAAAG
CAGCCTAGGACGGAAATACAAAAAGCAAGACGGATTGCCGGGCTTAAAATATATGGAGAGTCAGACAGAGT
TGGGCAACCGGCTGCAAGTTCGATCACGTGTTCAGGATTAAAGAGACGCGCGCTATTTCGACGCGAACGTT
CAGCAGTTTCTGTATTTCGCTCGAAATCCGCCAACGTGTTTAGGCAATATTTTTCCCACCTATTACAACAT
ACTAACGGCAACAGAGTTGAATGAATCGGACCAAGGCAACTGAATCTTCGGGATACCCGTTCTGGAAGGTT
TTCCAAGGTGGTTCACCGGATTCCACGCAATTAATCTGAACCTCTGTCAGCGACGTGGACCGGTCATATAA
TCCGAGAGGCTGGTGTAGCGAGAGTGCCGCGTAGGGAGGCAGGAGAGAGGGCGAGCGATAAAATTAATGGG
ATATCTAGGGTATCTAATCGAAGAGCATTTCAAGGCCTGGCCTTAAGGTGGCCACGTAAAGCTTAGCTTAA
TTGGGCAGCGAGTGGTGGCCGCGCCTTATTAGCGGGTTACAACATCCGGCTAGCTGTTCATCTGAAAAGTC
CTGCATGTATCCTGCGGCGGGTCTGAAAGACCAGTCGGTCCGCAGAGAAGCGACAGAGAAAGAGAGGGCAA
```

```
AGAGAGAGAAAGAGAGAAGCGGAGCAACTAGCCTATAATAAGCCGTACGGTTGTGTACGTTCTGTTCTGTT
CCGTTCCGTTCCTTCATGCTGCGCCAGTGAAGGCAGACCCATAAATCTCTCGGGCATACTCCAGGGGACAT
TTGATCCTTGCGTAACCCAACGCCGTTAATTTATTGCCCCGCGTTTAAAGTCACCTCCGTCCGCGAGAGAG
CAACGGCTACGCCGCCATCAGAACGCCGGTTCAGCATAAACCGACGGTTGTGGCTAGCTTTTCTTGTTATC
GACCTCTTTAACACGCCCTACACTTTGCTTCGTTCGCAACATCCATCCTTCGATCGAGAAAAAATTCCTGT
TCGCGGCGTCGTTATATCGGAATCGAGGTTGCCCACACCTACGGAGACCTTGCGATACAGCGTTGTTTCAG
CTGTTCGCTGTTCTTCCTTCTCGCGGTATTCCATTAGCAAAGAAAATCTCACACCCTCGCTATCTACATCG
TGTTTCCCACGACGCGTTCCTACAATGAAAACGCCATTAATTTCTTTTCAGCCGGCCAGAATGTATACCGG
TGATTAATTCGCCGGTGATGTTGCGCCCTGATAACGCTTTCACACCGTTTTCCAGGCACTCGGTTGCACCA
ATACAATTCAAGCATTTAGTATCGCACCCTACGTTTCATACGCTCGTTCGAAGAAGAGGAAAAAGCAAAGG
AAATGTTCATCGAACATTTTACAAGATAATTCGACTGTAGACTGTCGTGGCTTCTATCTGTTCGCTAAAAC
TACTGTTTAGGTTCTTATATCGTTTGCTGTAATTTGTCACGTGGTCGCTATGAACGTGTATTCTTATCCTT
AGCACGTGAACTTCCTCTTGTGTTTCACTCCAACGTATTTCCCTTGTCCTTGCAGGTCGAAAATTCGAATA
GCTCGTTTTCCATAAGTTACAAGTTTCCATAAGTTTACAAGGCGATTAAAGTTTCTGTTCACGTTGGCCAT
CGGACGTCTGTTTAATTAAGCTCGCATGTTGGACGCATGACGAGAAAAAGCCCGACGTCTAATTCAGCGGG
CGGACAGGGCCTAATCAAGATTCGAGTCCTCGCCGCTGGAATCGATCGTCTTCGAAACAGACAGCTGCATA
CGTAAGTCGAGAGGATATTGCCAATGTCCTTGCGCTGTCTTGTTGTTAGGATAATAGTTATCGTTCGTTTC
TGATTGTAGCATTTGCATGGCATTATAATATCGTCGCTATCTGCTTTACTATGAATACTCTTCAATTTCCA
ATGCAATCCACGCTAATGACAAATTATTCTTGAACCGAGAGAAGATCCTCTTCTACTATTTGGCGTTCTTC
GTTATAGAAATCTGATTGGTCGATCTACGTCGTATGTATTGCCCTCTATGGCATACGCTAGCCACGTGAAC
GTTGAGGATTCTCTTGCAATTAGGCAATGAGAGTGTTATCGATAAAGAGGCACGAACTAATCGGCACGGAG
CAGCTCGAGCGTGACTCGACACCGTTTCATTATTCCGTGAACGCTTTGATGATCTTCTGGGTTAGACGGCG
TACACCGTCTTCAAGAGGCACCTTCATTAACCGGCTGACGAATGAACATTAATAAGCCGGGCTTCCAAGTG
GAGGCGCGTTCGCGCGCGACTCATCAGCCTCTTTGATGAAACCCGAGATCTATGTTTATCCCGGGCATCCT
TTGATTGGCTGGCTAGGACTCCAATTTTGCGGTTTTGAACACGGCACGGCGTACAAACTCGTCGCTGATTG
CCCCAGCGTCTGTAAATTGGGCATCAAAGAATGTAGGTAATATTAGAAAGTTTAATTCTAATGTATTTTTA
AAAGCGCAGTTTACTCCGGCCGCTGAAGGATTCCAGCAAAAACTGTAGGCTGTGATGCTGAGTAATTGTTC
GCCAACAGGCAGCGAATCAGAGGAAGCTAGGCGAAATGGTTAAAAGTTATAAGAAAACGCTAGAGATTGGT
ATGTTGTTAATTTGAGAGCGTAGCCCGAGGGAGAGAGAAAGAAAAGGAGGAGAGGGCATGTTCCCAGAGTT
TCTCTCCTTCCCTTGGCGCGCTGCTCTTCGTCCGGTCGAGCAACGCTTGGAACCCTCCAAGAGAAGAGAAG
AGAATCTCCTCTTTCTCTCTCGCTGTGGAGGAGAGTCGCCATCCGTCTTCATCGGCCGACAGCGCGACGGG
CGGCGTTTGCCTCGAATGCTCACCAAGGAGTTTTCTTCTTCTTCTTCTACTCCTCCTCGTACTCCACCTCC
TCCTCCTCCTCTTCCTTCTTCTCGCTGCTGCTCATCCCTTCGCCCCCTCTTCTTTTTTCTCTCCATCGTAC
CACTCTGAAATATTCTCCTTTTGCCCTCTTGCACTCGTGTGGCAACAGCTAACCTAACCCTTTGAGATTAT
GTCTGGTGAGAGAGTATGCATACACATATACCTGGCCACCGCGCAAGTACCAATATAAACACGGACACGCG
GTCCCAGACGTTGCTTCCATACACAGAGTCCAGCTGGTTTTCTGTGTACGTCCGCGCTTCTCCCTTTCCCT
TGCTCCCTTGGCAAGTACAATTAATATTCTTGTTCGGGATTCGTGTACCGCCACACCGAGCGCAGATCATG
TCACGGATATCTGTAAGGTATCTATCATACGCGGAACTGCGCCTACGAACCGGCGGCCGAGAGCTTTTTCT
TCCTTGGACAGACATTTTTGTTTCGTCCCGAGCCCGTAAAAACTTTCCCGCGTGCCGAAGCAGGTTGCAGG
CCGTGGCGTTCATTAAATATCTCAGATTGCGGCCTCGAGCTCTCCCTGCCGAGTATACGCGCCTCTAACCA
AGCCTCCGTGTCTCCCATCATCCTCTTTCCCTATGCCATGGCTATTTCCAATGGATCGGGATGTCGCTGAA
CGCCAGAATTCGGCAATTCTTCTATTGTCTTTCTTTTTCTTGAGTTTTTTTTTAGGATTACACTGGAACAA
ATGTCGACAGTACAGAAGTTATTCTTATACCGAGTTGGAATCGTTCTGAAGAGGTTAGGATCAGACTGTAT
AGTTATCAAGTTACTCGGTAGCAGGATAATTAAGAAACAAAACTATTTCCAGTGATCGGATCACATTGGTA
TTGGCTTACAGTGGTCTTCTTACGTTTTAGCGAGAAACCAAGTTTCTTGGAACGACGTCGAAATTAGAGCG
CGCATAATGTTCGAGTATTATGCGAGCAGTTTTGAAACCACAGTACGTTTCAAGGGTAATTCGCGTGCACA
TGCGTCCGCGCAGTGTCTGTGAAATAAAGAGGAGGAATAAAAAAGAAGAACAAGTTACGGGGGTTACATGT
AACGTAGGATCGATCGTAGGATCAAGGACGCGAGTTGGTCTACATCGAACCTGCTCGTAAGAGACCTGGCG
GTGAAGGAAACGGACGCGTAGCCTCGATTCACGCCAGATCTCTGACGAGTATATATTCGTAGCCAGGGAAA
AAGGAGGAGAAACGCGAGACTAACGAGGAATCCGCGCGGTGATTTCCATGTTTCAGAGAGGAAACGAGGCC
GGCAAACGTACACCCGATACCAAACCCTCGAGCTGGAGAAGGAGTTCCACTTCAACCGATACCTGACCAGG
CGGCGGCGCATCGAGATCGCGCACGCACTCTGCCTGACGGAACGGCAAATCAAAATCTGGTTCCAAAACAG
ACGGATGAAATGGAAGAAGGAGAACAAGACGAAGGGCGAACCGGGCTCGGGCGACGGCGACACTGAAATCT
CGCCGCAGACATCGCCGCAGGGTTGA
```

Coding Sequence

```
>antp_nv_cds_1089_bp
ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGATCTGCGCAATGGCGGGGTGGAGCATCCGCATCAGCA
CCAGCAGCACTACGGCGCGGCGGTCCAGGTGCCCCAGCAGCAGCAGGCCGTGCAGCAGCAGCCCCAGCAGG
CGAGCGACCCCTGCGACCCGTCGATGCTGCGCCAAGGCGTGCCGGGCCACCACGGCTACGGGGCCGCGACG
GGCCAGCAGCCGGGCATGCCCTACCCCCGCTTCCCGCCCTACGACCGCATGGACATCAGGAACGCGGCCTA
CTACCAGCAGCAGCAGCAGGAGCACGGCATGGACATGGCCAGCTACCGGGCGAGCTCGCCGAGCGCGGGCA
TGGCCGGCCTCCACATGGGCCACACGCCGACCCCGGTCAACGGCCACCCCGCCAGCACGCCCATCGTCTAC
GCGAGCTGCAAGCTCCAAGCGGCGGCGGTCGACCACCAGGGCAGCGTCCTCGACGGGCCCGATAGTCCGCC
GCTCGTCGACGCCCAGATGCACCACCAGATGCACCCCCAGCACACGCACATGCAGGCCCAGCAGTCGCACC
CCCAGCAGCAGCCCCAGCCTCAAGCGCCTCACCAGCAGGCCCACATGCAACCCCAGCAGACGCAGCAGCAG
CACATGATGTACCAGCAGCAGACGCAGCCCCAGCAGCCCCAGCCCGCGGCGATGCACCCCCAGCAGCAGGC
CCAGCAGCAGCAGCACCAGGGCGTCGTCGCCTCGCCGCTCGGCCAGCAGCAGCCCGGCACGCCCCAGAGCG
CCGCGCCGACGAACCTGCCCAGTCCCCTCTACCCCTGGATGAGGAGTCAGTTTGAGAGGAAGCGTGGCCGG
CAGACGTACACGCGATACCAGACCCTCGAGCTCGAGAAGGAGTTCCACTTCAACCGCTACCTGACCAGGCG
ACGCAGGATCGAGATCGCGCATGCGCTCTGCCTGACCGAGCGCCAGATCAAGATCTGGTTCCAGAACAGGC
GCATGAAGTGGAAAAAGGAGACCAAGACGAAGGGCGAGCCGAACTCGGGAGACGGTGACACCGACATCTCG
CCCCAGACCTCGCCCCAGGGTTGA
```

```
>antp_bt_cDNA_1059_bp
ATGAGTTCGTACTTCGCGAATTCGTACATCCCGGACCTGCGTAATGGCGGGGTGGAACACCCGCATCAGCA
TCAGCAGCACTACGGTGCCGCCGTCCAGGTGCCCCAGCAAACGCAGTCGGTACAGCAGCCATCTCAGCAAA
CCGGGGATCCATGCGATCCTAGTCTCCTACGTCAGGGAGTGCCTGGCCATCACTATGGGGCCGCTGGTAGC
CAGCAAGATATGCCTTATCCGAGGTTTCCTCCGTACAATCGGATGGACATGCGGAACGCGACCTATTATCA
GCATCAACAGGAGCACGGCAGCATGGACGGGTTGGGTGGTTACAGGTCGACGTCCCCGAGCCCCGGTATGG
GCCACATGGGACACACACCGACCCCGAACGGACATCCGTCCACTCCTATTGTCTATGCGAGTTGCAAGCTT
CAAGCGGCCGCGGTCGATCATCAGGGTAGCGTACTCGATGGACCGGACAGCCCGCCATTGGTCGAGTCGCA
GATGCACCACCAAATGCATTCGCAACATCCTCACATGCAGCCGCAACAGTCACAACATCAACAGCAGCAGC
AGCAGCATCAACATCTTCAGGCGCAGCAGCAGCACATGATGTACCAACAGCAACAACAAACACAGGCGGCG
TCGCAACAATCTCAGCCTGGCATGCATCCGCAACAACAACAGCAACCTCAGCAACACCAAGGGGTGGTCAC
GTCGCCGCTTAGTCAGCAACAGCAGGCCGCTCCTCAAGGTGCGGCCACTGCCAACCTACCAAGTCCGCTCT
ACCCGTGGATGAGAAGTCAATTCGAGAGGAAACGAGGCCGGCAAACGTACACCCGATACCAAACCCTCGAG
CTGGAGAAGGAGTTCCACTTCAACCGATACCTGACCAGGCGGCGGCGCATCGAGATCGCGCACGCACTCTG
CCTGACGGAACGGCAAATCAAAATCTGGTTCCAAAACAGACGGATGAAATGGAAGAAGGAGAACAAGACGA
AGGGCGAACCGGGCTCGGGCGACGGCGACACTGAAATCTCGCCGCAGACATCGCCGCAGGGTTGA
```

Deduced peptides

```
>antp_nv_aa_362_residues
MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQQQAVQQQPQQASDPCDPSMLRQGVPGHHGYGAAT
GQQPGMPYPRFPPYDRMDIRNAAYYQQQQQEHGMDMASYRASSPSAGMAGLHMGHTPTPVNGHPASTPIVY
ASCKLQAAAVDHQGSVLDGPDSPPLVDAQMHHQMHPQHTHMQAQQSHPQQQPQPQAPHQQAHMQPQQTQQQ
HMMYQQQTQPQQPQPAAMHPQQQAQQQQHQGVVASPLGQQQPGTPQSAAPTNLPSPLYPWMRSQFERKRGR
QTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKETKTKGEPNSGDGDTDIS
PQTSPQG
```

```
>antp_bt_aa_352_residues
MSSYFANSYIPDLRNGGVEHPHQHQQHYGAAVQVPQQTQSVQQPSQQTGDPCDPSLLRQGVPGHHYGAAGS
QQDMPYPRFPPYNRMDMRNATYYQHQQEHGSMDGLGGYRSTSPSPGMGHMGHTPTPNGHPSTPIVYASCKL
QAAAVDHQGSVLDGPDSPPLVESQMHHQMHSQHPHMQPQQSQHQQQQQHQHLQAQQQHMMYQQQQTQAA
```

```
SQQSQPGMHPQQQQQPQQHQGVVTSPLSQQQQAAPQGAATANLPSPLYPWMRSQFERKRGRQTYTRYQTLE
LEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSGDGDTEISPQTSPQG
```

Supplementary data C-2. Positions (bp) with multiple codon changes along the nucleotide alignment of the coding sequence for the *Antennapedia* gene from *Apis mellifera*, *Bombus terrestris* and *Nasonia vitripennis*.

133, 145, 166, 211, 214, 256, 271, 289, 292, 295, 298, 319, 355, 505, 514, 538, 547, 568, 619, 667, 673, 685, 691, 730, 766, 784, 793, 802, 805, 943, 1045.

Table C-1.Summary of synonymous and non-synonymous changes along branch 1 (C-1.1) in the gene phylogeny reconstructed for *Antennapedia* (*Antp*). We used coding nucleotide sequences from three species of Hymenoptera Apis mellifera (am), Bombus terrestris (Bt) and Nasonia vitripennis (nv) to perform an empirical Bayes reconstruction of ancestral sequences. Data were obtained using the feature `Rate_Ancestor` in the program `codeml` from the PAML4 (Zang, 2007) package. *p* is the posterior probability. The columns 'Polarity' and 'Charge' are labeled 'yes' (y) or 'no (n) according to the effect of the observed nonsynonymous substitution in the ionic charge and polarity at each position. Shaded rows highlight the position of nonsynonymous substitutions. Check root of tree (Figure 4.1) for directions of change. (C-1.2. contains data for branch 2, C-1.3 for branch 3 and C-1.4 for branch 4).

Table C-1.1.

Branch 1	4..5							Polarity	Charge
Position	node 4		*p*		node 5		*p*		
14	CGC	(R)	0.998	->	CGT	(R)	1		
19	GAG	(E)	0.998	->	GAA	(E)	1		
22	CAC	(H)	0.997	->	CAT	(H)	1		
24	CAC	(H)	0.996	->	CAT	(H)	1		
29	GGC	(G)	1	->	GGT	(G)	1		
31	GCG	(A)	0.999	->	GCC	(A)	1		
38	AAG	(T)	0.242	->	ACG	(T)	0.998		
40	GCG	(A)	0.529	->	TCG	(S)	0.999	y	y
48	GCG	(A)	0.97	->	GCC	(A)	0.947		
49	GGC	(G)	0.696	->	GGG	(G)	1		
50	GAC	(D)	0.998	->	GAT	(D)	1		
53	GAC	(D)	0.998	->	GAT	(D)	1		
55	ACC	(T)	0.476	->	AGC	(S)	0.999	n	n
56	ATG	(L)	0.538	->	CTG	(L)	0.999		
57	CTG	(L)	0.999	->	CTA	(L)	1		
58	CGC	(R)	0.998	->	CGT	(R)	1		
62	CCG	(P)	0.998	->	CCT	(P)	1		
70	GGG	(G)	0.351	->	GGC	(G)	0.999		
73	CAG	(Q)	0.995	->	CAA	(Q)	1		
76	CCC	(P)	0.998	->	CCT	(P)	1		
77	TAC	(Y)	0.999	->	TAT	(Y)	1		
79	CGG	(R)	0.998	->	AGG	(R)	1		
82	CCC	(P)	0.998	->	CCG	(P)	1		
85	CGC	(R)	0.998	->	CGG	(R)	1		
93	TAC	(Y)	0.999	->	TAT	(Y)	1		
94	TAC	(Y)	0.999	->	TAT	(Y)	1		
96	CAG	(H)	0.518	->	CAC	(H)	0.997		
97	CAG	(Q)	0.995	->	CAA	(Q)	1		
100	CAG	(Q)	0.349	->	CAC	(H)	0.995	n	y
101	GAC	(D)	0.267	->	GGC	(G)	0.999	n	y
106	GGC	(G)	0.627	->	GGT	(G)	1		
107	GGC	(G)	0.609	->	GGT	(G)	1		
111	ACC	(T)	0.675	->	ACG	(T)	0.97		
115	CCG	(P)	0.574	->	CCC	(P)	0.999		
121	GGC	(G)	0.999	->	GGA	(G)	1		
128	GGC	(G)	0.999	->	GGA	(G)	1		
131	ACC	(S)	0.497	->	TCC	(S)	0.999		
132	ACG	(T)	0.999	->	ACT	(T)	1		
133	CCC	(P)	0.998	->	CCT	(P)	1		
141	CTC	(L)	0.999	->	CTT	(L)	1		
147	GAC	(D)	0.998	->	GAT	(D)	1		
150	GGC	(G)	0.999	->	GGT	(G)	1		
152	GTC	(V)	0.999	->	GTA	(V)	1		
154	GAC	(D)	0.998	->	GAT	(D)	1		
164	GCG	(A)	0.529	->	TCG	(S)	0.999	y	y
169	CAG	(Q)	0.995	->	CAA	(Q)	1		
172	CCG	(P)	0.583	->	TCG	(S)	0.638	y	y
173	CAG	(Q)	0.995	->	CAA	(Q)	1		
175	CCG	(P)	0.584	->	CCC	(P)	0.999		
179	GCC	(P)	0.504	->	CCG	(P)	0.998		
182	TCG	(S)	0.992	->	TCC	(S)	0.988		
188	GCC	(A)	0.984	->	GCG	(A)	0.967		
190	AAG	(Q)	0.237	->	CAG	(Q)	0.852		
192	CAC	(H)	0.648	->	CAT	(H)	1		
193	CTG	(L)	0.668	->	CTT	(L)	1		
208	CCG	(P)	0.544	->	ACG	(T)	0.559	y	y
211	CCG	(P)	0.557	->	CAG	(Q)	0.998	y	y
212	CCG	(P)	0.422	->	TCG	(S)	0.999	y	y

Table C-1.2.

Branch 2	*5..1*						*Polarity*	*Charge*
Position	*node 5*		p		*am_antp*			
4	TAC	(Y)	1	->	TAT	(Y)		
43	CAG	(Q)	1	->	CAA	(Q)		
50	GAT	(D)	1	->	GAC	(D)		
53	GAT	(D)	1	->	GAC	(D)		
58	CGT	(R)	1	->	CGC	(R)		
62	CCT	(P)	1	->	CCC	(P)		
65	CAC	(H)	1	->	CAT	(H)		
68	GCC	(A)	1	->	GCG	(A)		
82	CCG	(P)	1	->	CCC	(P)		
89	CGG	(R)	1	->	CGT	(R)		
92	ACC	(T)	0.999	->	ACG	(T)		
99	GAG	(E)	0.951	->	GAC	(D)	n	y
101	GGC	(G)	0.999	->	GGG	(G)		
111	ACG	(T)	0.97	->	GCG	(A)	y	y
115	CCC	(P)	0.999	->	CCT	(P)		
121	GGA	(G)	1	->	GGG	(G)		
126	CCG	(P)	1	->	CCT	(P)		
128	GGA	(G)	1	->	GGG	(G)		
132	ACT	(T)	1	->	ACC	(T)		
133	CCT	(P)	1	->	CCG	(P)		
135	GTC	(V)	1	->	GTG	(V)		
141	CTT	(L)	1	->	CTG	(L)		
150	GGT	(G)	1	->	GGG	(G)		
152	GTA	(V)	1	->	GTG	(V)		
154	GAT	(D)	1	->	GAC	(D)		
172	TCG	(S)	0.638	->	ACG	(T)	n	n
182	TCC	(S)	0.988	->	GGC	(G)	n	n
186	CAG	(Q)	0.924	->	TCG	(S)	n	n
187	CAG	(Q)	1	->	CAA	(Q)		
188	GCG	(A)	0.967	->	GCA	(A)		
196	CAG	(Q)	0.929	->	CAC	(H)	n	y
197	CAG	(Q)	0.94	->	GAG	(E)	n	y
208	ACG	(T)	0.559	->	TCG	(S)	n	n
214	CCT	(P)	1	->	CCA	(P)		
219	CAA	(Q)	0.948	->	CGA	(R)	n	y
224	CCT	(P)	0.972	->	GCT	(A)	n	n
235	CTT	(L)	1	->	CTA	(L)		
239	CAG	(Q)	1	->	CAA	(Q)		
247	GCC	(A)	1	->	GCA	(A)		
248	ACC	(T)	0.99	->	AGC	(S)	n	n
253	AGT	(S)	1	->	AGC	(S)		
254	CCG	(P)	1	->	CCT	(P)		
255	CTC	(L)	1	->	CTG	(L)		
286	TTC	(F)	0.995	->	TAC	(Y)	y	y
294	CGG	(R)	1	->	CGT	(R)		
301	GCA	(A)	1	->	GCC	(A)		
304	CTG	(L)	1	->	CTT	(L)		
313	TTC	(F)	1	->	TTT	(F)		
326	ACG	(T)	0.977	->	TCG	(S)	n	n
329	GAG	(E)	0.957	->	ACG	(T)	n	y
330	CCG	(P)	1	->	CCC	(P)		
335	GGC	(G)	1	->	GGG	(G)		
338	GAA	(E)	0.997	->	GAG	(E)		
343	ACA	(T)	1	->	ACG	(T)		

Table C-1.3.

Branch 3	*5..2*					*Polarity*	*Charge*
Position	*node 5*		p	*bt_antp*			
30	GCG	(A)	1 ->	GCC	(A)		
37	CAG	(Q)	1 ->	CAA	(Q)		
41	GTG	(V)	1 ->	GTA	(V)		
44	CAG	(Q)	0.9 ->	CCA	(P)	y	y
45	TCC	(S)	1 ->	TCT	(S)		
47	CAG	(Q)	1 ->	CAA	(Q)		
48	GCC	(A)	0.947 ->	ACC	(T)	y	y
51	CCG	(P)	1 ->	CCA	(P)		
54	CCG	(P)	1 ->	CCT	(P)		
55	AGC	(S)	0.999 ->	AGT	(S)		
56	CTG	(L)	0.999 ->	CTC	(L)		
60	GGC	(G)	1 ->	GGA	(G)		
64	CAC	(H)	1 ->	CAT	(H)		
66	TAC	(Y)	1 ->	TAT	(Y)		
69	GCG	(A)	1 ->	GCT	(A)		
70	GGC	(G)	0.999 ->	GGT	(G)		
74	GAC	(D)	0.999 ->	GAT	(D)		
80	TTC	(F)	1 ->	TTT	(F)		
81	CCG	(P)	1 ->	CCT	(P)		
84	AAC	(N)	0.999 ->	AAT	(N)		
96	CAC	(H)	0.997 ->	CAT	(H)		
105	ATG	(M)	0.982 ->	TTG	(L)	n	n
112	TCG	(S)	1 ->	TCC	(S)		
116	GGC	(G)	1 ->	GGT	(G)		
123	ACG	(T)	1 ->	ACA	(T)		
129	CAC	(H)	1 ->	CAT	(H)		
134	ATC	(I)	1 ->	ATT	(I)		
136	TAC	(Y)	1 ->	TAT	(Y)		
138	AGC	(S)	1 ->	AGT	(S)		
142	CAG	(Q)	1 ->	CAA	(Q)		
144	GCG	(A)	1 ->	GCC	(A)		
148	CAC	(H)	1 ->	CAT	(H)		
155	GGG	(G)	1 ->	GGA	(G)		
160	CCG	(P)	1 ->	CCA	(P)		
161	CTG	(L)	1 ->	TTG	(L)		
171	CAC	(H)	1 ->	CAT	(H)		
174	CAC	(H)	1 ->	CAT	(H)		
175	CCC	(P)	0.999 ->	CCT	(P)		
180	CAG	(Q)	1 ->	CAA	(Q)		
182	TCC	(S)	0.988 ->	TCA	(S)		
183	CAG	(Q)	0.996 ->	CAA	(Q)		
184	CAC	(H)	0.999 ->	CAT	(H)		
185	CAG	(Q)	1 ->	CAA	(Q)		
188	GCG	(A)	0.967 ->	CAG	(Q)	y	y
190	CAG	(Q)	0.852 ->	CAT	(H)	n	y
191	CAG	(Q)	1 ->	CAA	(Q)		
203	CAG	(Q)	1 ->	CAA	(Q)		
205	CAG	(Q)	1 ->	CAA	(Q)		
206	CAG	(Q)	0.997 ->	CAA	(Q)		
207	CAG	(Q)	1 ->	CAA	(Q)		
208	ACG	(T)	0.559 ->	ACA	(T)		
210	CAG	(Q)	1 ->	CAA	(Q)		
211	CAG	(Q)	0.998 ->	CAA	(Q)		
212	TCG	(S)	0.999 ->	TCT	(S)		
217	CAC	(H)	1 ->	CAT	(H)		
220	CAG	(Q)	1 ->	CAA	(Q)		
221	CAG	(Q)	1 ->	CAA	(Q)		
236	AGC	(S)	1 ->	AGT	(S)		

Table C-1.4.

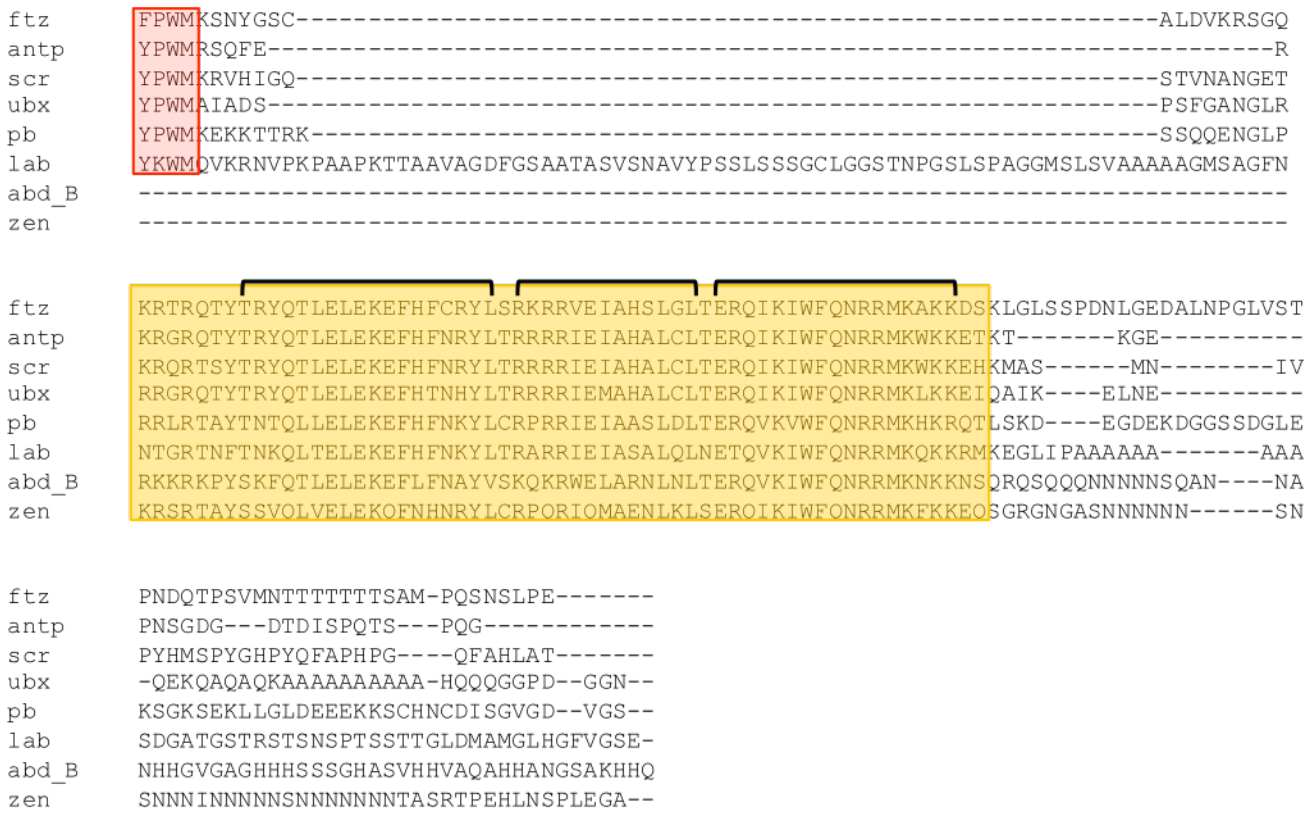
Branch 4	*4..3*							
	node 4		p		*nv_antp*		*Polarity*	*Charge*
12	GAC	(D)	0.998	->	GAT	(D)		
20	CAC	(H)	0.996	->	CAT	(H)		
22	CAC	(H)	0.997	->	CAT	(H)		
38	AAG	(T)	0.242	->	CAG	(Q)	n	n
40	GCG	(A)	0.529	->	GCC	(A)		
45	TCC	(S)	0.596	->	CCC	(P)	y	y
49	GGC	(G)	0.696	->	AGC	(S)	n	n
51	CCG	(P)	0.998	->	CCC	(P)		
55	ACC	(T)	0.476	->	TCG	(S)	n	n
56	L		0.538	->	M		n	n
59	CAG	(Q)	0.995	->	CAA	(Q)		
70	GGG	(G)	0.351	->	ACG	(T)	n	n
71	AGC	(S)	0.581	->	GGC	(G)	n	n
74	GAC	(D)	0.616	->	GGC	(G)	n	y
78	CCG	(P)	0.998	->	CCC	(P)		
79	CGG	(R)	0.998	->	CGC	(R)		
84	AAC	(N)	0.565	->	GAC	(D)		
88	ATG	(M)	0.504	->	ATC	(I)	n	n
89	CGG	(R)	0.998	->	AGG	(R)		
92	ACC	(T)	0.512	->	GCC	(A)	y	y
96	H		0.518	->	Q		n	y
99	GAG	(E)	0.502	->	CAG	(Q)	n	y
100	CAG	(Q)	0.349	->	GAG	(E)	n	y
101	GAC	(D)	0.267	->	CAC	(H)	n	y
102	AGC	(S)	0.581	->	GGC	(G)	n	n
106	GGC	(G)	0.627	->	GCC	(A)	y	y
107	GGC	(G)	0.609	->	AGC	(S)	n	n
109	AGG	(R)	0.998	->	CGG	(R)		
110	TCG	(S)	0.537	->	GCG	(A)	y	y
111	ACC	(T)	0.675	->	AGC	(S)	n	n
115	CCG	(P)	0.574	->	GCG	(A)	n	n
118	GGC	(G)	0.527	->	GCC	(A)	y	y
130	CCG	(P)	0.998	->	CCC	(P)		
131	ACC	(S)	0.497	->	AGC	(S)		
142	CAG	(Q)	0.996	->	CAA	(Q)		
156	CCG	(P)	0.998	->	CCC	(P)		
157	GAC	(D)	0.998	->	GAT	(D)		
158	AGC	(S)	1	->	AGT	(S)		
161	CTG	(L)	0.998	->	CTC	(L)		
163	GAG	(E)	0.515	->	GAC	(D)	n	n
164	GCG	(A)	0.529	->	GCC	(A)		
172	CCG	(P)	0.583	->	CCC	(P)		
175	CCG	(P)	0.584	->	ACG	(T)	y	y
179	P		0.504	->	A		n	n
183	CAG	(Q)	0.516	->	CAC	(H)	n	y
184	CAC	(H)	0.613	->	CCC	(P)	y	y
189	CAG	(Q)	0.515	->	CAC	(H)	n	y

Table C-1.4. Continued.

Branch 4	*4..3*							
	node 4		p		*nv_antp*		*Polarity*	*Charge*
190	AAG	(Q)	0.237	->	ATG	(M)	y	y
191	CAG	(Q)	0.996	->	CAA	(Q)		
192	CAC	(H)	0.648	->	CCC	(P)	y	y
193	CTG	(L)	0.668	->	CAG	(Q)	y	y
195	GCG	(A)	0.515	->	ACG	(T)	y	y
206	CAG	(Q)	0.317	->	ACG	(T)	n	n
208	CCG	(P)	0.544	->	CCC	(P)		
211	CCG	(P)	0.557	->	CCC	(P)		
212	CCG	(P)	0.422	->	CAG	(Q)	y	y
213	CCG	(P)	0.587	->	CCC	(P)		
214	CCG	(P)	0.586	->	GCG	(A)	n	n
215	GGC	(G)	0.526	->	GCG	(A)	y	y
218	CCG	(P)	0.998	->	CCC	(P)		
222	CCG	(P)	0.327	->	GCC	(A)	n	n
224	CCG	(P)	0.632	->	CAG	(Q)	y	y
229	GGG	(G)	0.999	->	GGC	(G)		
230	GTG	(V)	0.999	->	GTC	(V)		
232	GCG	(A)	0.508	->	GCC	(A)		
236	AGC	(S)	0.618	->	GGC	(G)	n	n
240	CCG	(P)	0.587	->	CCC	(P)		
241	A		0.494	->	G		y	y
242	GCG	(A)	0.568	->	ACG	(T)	y	y
245	G		0.511	->	S		n	n
246	GCG	(A)	0.999	->	GCC	(A)		
248	ACG	(T)	0.559	->	CCG	(P)		
249	GCG	(A)	0.552	->	ACG	(T)	y	y
252	CCG	(P)	0.998	->	CCC	(P)		
257	CCG	(P)	0.998	->	CCC	(P)		
260	AGA	(R)	0.999	->	AGG	(R)		
263	TTC	(F)	1	->	TTT	(F)		
267	CGG	(R)	0.999	->	CGT	(R)		
293	CGG	(R)	0.998	->	CGA	(R)		
295	CGG	(R)	0.998	->	AGG	(R)		
300	CAC	(H)	0.996	->	CAT	(H)		
316	AGA	(R)	0.999	->	AGG	(R)		
321	AAG	(K)	0.998	->	AAA	(K)		
324	AAC	(N)	0.571	->	ACC	(T)	n	n
331	GAC	(D)	0.353	->	AAC	(N)	n	y
333	GGC	(G)	1	->	GGA	(G)		
338	E		0.519	->	D		n	n
341	CCG	(P)	0.998	->	CCC	(P)		
345	CCG	(P)	0.998	->	CCC	(P)		

Figure C-1. Multiple sequence alignment of the homeodomain of *Hox* genes in *Nasonia vitripennis*. 1) Amino acid sequence alignment of eight of the ten genes in the *N. vitripennis* Hox cluster. 2) Sequence logo of the *Hox* gene homeodomain and surrounding amino acids in *N. vitripennis*. A high degree of conservation is apparent across the homeodomain of all members of the *Hox* family in *N. vitripennis* and at two other places located 5’ of the homeodomain. The YPWM motif is highlighted in red. One other conserved run of approximately 30 nucleotide residues is located upstream of the homeodomain (not shown here). The three helices formed by the amino acids in the homeodomain are marked with black brackets.

Figure C-1.1.

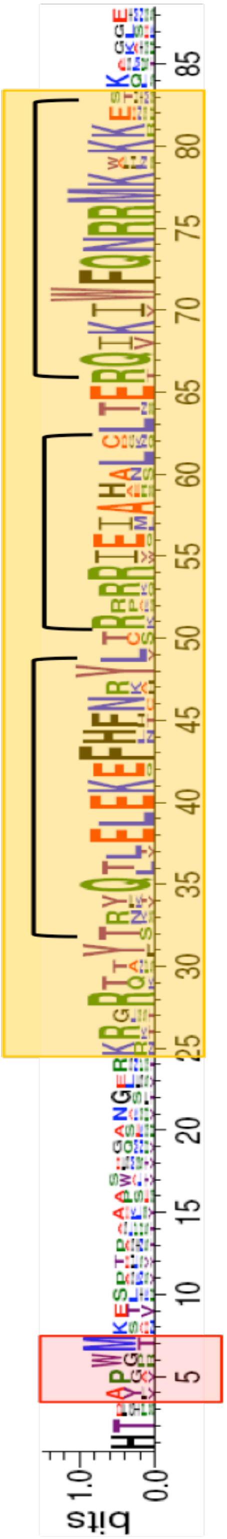


```
ftz      FPWMKSNYGSC------------------------------------------------------------ALDVKRSGQ
antp     YPWMRSQFE----------------------------------------------------------------------R
scr      YPWMKRVHIGQ------------------------------------------------------------STVNANGET
ubx      YPWMAIADS--------------------------------------------------------------PSFGANGLR
pb       YPWMKEKKTTRK-----------------------------------------------------------SSQQENGLP
lab      YKWMQVKRNVPKPAAPKTTAAVAGDFGSAATASVSNAVYPSSLSSSGCLGGSTNPGSLSPAGGMSLSVAAAAAGMSAGFN
abd_B    --------------------------------------------------------------------------------
zen      --------------------------------------------------------------------------------

ftz      KRTRQTYTRYQTLELEKEFHFCRYLSRKRRVEIAHSLGLTERQIKIWFQNRRMKAKKDSKLGLSSPDNLGEDALNPGLVST
antp     KRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKETKT------KGE-----------
scr      KRQRTSYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEHKMAS------MN--------IV
ubx      RRGRQTYTRYQTLELEKEFHTNHYLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEIQAIK----ELNE----------
pb       RRLRTAYTNTQLLELEKEFHFNKYLCRPRRIEIAASLDLTERQVKVWFQNRRMKHKRQTLSKD----EGDEKDGGSSDGLE
lab      NTGRTNFTNKQLTELEKEFHFNKYLTRARRIEIASALQLNETQVKIWFQNRRMKQKKRMKEGLIPAAAAAA-------AAA
abd_B    RKKRKPYSKFQTLELEKEFLFNAYVSKQKRWELARNLNLTERQVKIWFQNRRMKNKKNSQRQSQQQNNNNNSQAN----NA
zen      KRSRTAYSSVQLVELEKQFNHNRYLCRPQRIQMAENLKLSERQIKIWFQNRRMKFKKEQSGRGNGASNNNNNN------SN

ftz      PNDQTPSVMNTTTTTTTSAM-PQSNSLPE-------
antp     PNSGDG---DTDISPQTS---PQG------------
scr      PYHMSPYGHPYQFAPHPG----QFAHLAT-------
ubx      -QEKQAQAQKAAAAAAAAAA-HQQQGGPD--GGN--
pb       KSGKSEKLLGLDEEEKKSCHNCDISGVGD--VGS--
lab      SDGATGSTRSTSNSPTSSTTGLDMAMGLHGFVGSE-
abd_B    NHHGVGAGHHHSSSGHASVHHVAQAHHANGSAKHHQ
zen      SNNNINNNNNSNNNNNNNTASRTPEHLNSPLEGA--
```

Figure C-1.2.


bits