EVOLUTION AND EXPRESSION OF THE HIGHLY VARIABLE CELL ADHESION MOLECULE DSCAM IN THE CRUSTACEAN DAPHNIA AND OTHER ARTHROPODS

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät der
Universität Basel

von

Daniela Brites

Basel, 2012

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Fakultätsverantwortlicher: Prof. Dieter Ebert, Basel

Betreuer: Prof. Dieter Ebert, Basel

Emeritus Prof. Louis Du Pasquier, Basel

Externer Referent: Prof. Hinrich Schulenburg, Kiel

Basel, den 27 April 2010

Prof. Dr. Eberhard Parlow, Dekan

Dedico este trabalho a três fabulosas mulheres,

À minha mãe, Isabel

À minha avó Zaia

À minha tia Leopoldina

I dedicate my work to three great women,

My mother, Isabel

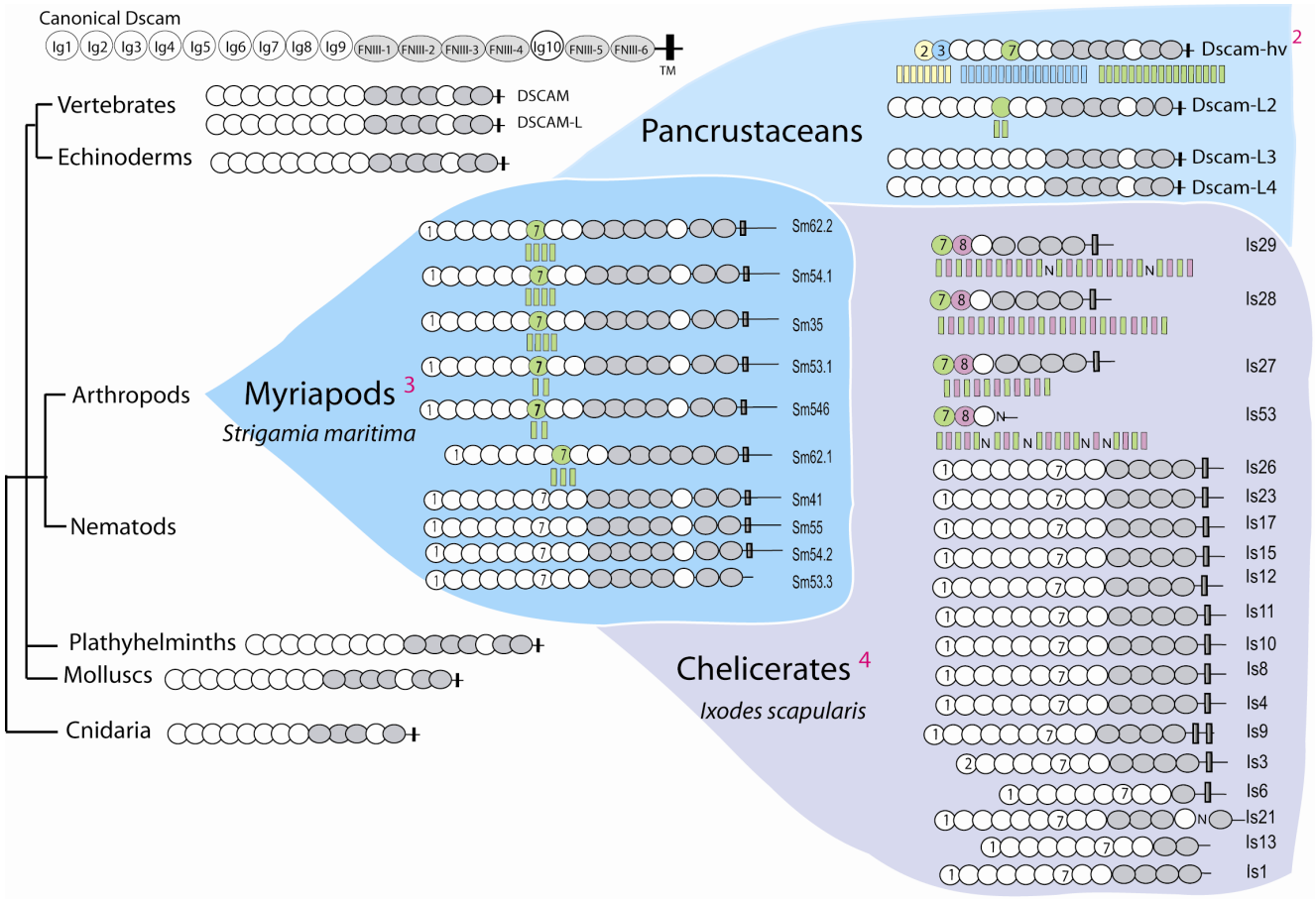
My grandmother Zaia

My aunt Leopoldina

Sábio é quem se contenta com o espectáculo do mundo[1]

Wise is he who enjoys the show offered by the world[1]

[1] 16.06.1914, Odes de Ricardo Reis, Fernado Pessoa



Down syndrome cell adhesion molecule (Dscam) reconstructions of different metazoa

TABLE OF CONTENTS

SUMMARY

The Down syndrome cell adhesion molecule (Dscam) family, is within the cell adhesion molecules, a family whose members are characterized by being composed of immunoglobulin (Ig) and fibronectin domains and which are known to play an essential role in the development of the nervous system in both vertebrates and invertebrates.

In insects, one member of the Dscam family diversified extensively due to internal exon duplications and a sophisticated mechanism of mutually exclusive alternative splicing (AS). This enables a single individual to generate somatically thousands of Dscam isoforms which differ in half of two Ig domains and in another complete Ig domain. That creates a high diversity of adhesion properties which are used by nervous cells and also by immune cells (hemocytes).

How this situation evolved is best understood my means of comparative studies. I have studied aspects of the evolution and expression of this diversified member of the Dscam family mainly in the brachiopod crustacean *Daphnia magna* and to lesser extent, in other representatives of the arthropod phyla. I have shown that like in insects, a highly variable Dscam gene evolved in crustaceans, which also express Dscam diversity in nervous and in immune cells. Additionally I could demonstrate that not only Dscam's ectodomains are diversified but that several cytoplasmic tails with different signal transduction capacities can also be expressed. The comparison between *Daphnia* and insects revealed furthermore that there is high amino acid conservation among distantly related species for most Dscam domains except for the Ig regions that are coded by the multiple exons, suggesting that the latter evolved under different selective constraints.

Dscam has been proposed as an exciting candidate molecule for mediating specific immune responses in arthropods. Nevertheless, the involvement of Dscam in immunity remains largely elusive. I tested the effect of parasite infection on the expression of total Dscam and on the diversity of some duplicated exons at the RNA level and found no significant effect. Yet, hemocytes expressed reduced transcript diversity relative to the brain, but each transcript was likely more abundant. This would be consistent with a function in the immune system given that each Dscam isoform would be present in higher concentrations which would increase their functional capacity.

Dscam isoforms engage in dimer formation with other identical isoforms, promoting cell-cell recognition. It has been demonstrated that the variable parts of Dscam coded by the duplicated exons mediate dimer formation. The genetic diversification caused by exon duplication and AS has thus direct functional implications. I estimated signatures of selection on some of the regions involved in dimer formation by comparing sequences from different *Daphnia magna* populations and from different species

of *Daphnia* and *Drosophila*. The results indicated that diversity created by duplication followed by divergence is maintained by purifying selection against new mutations and against new gene conversion events. That is consistent with the essential role of Dscam diversity in the nervous system. Contrastingly, I found that some parts of the variable regions which are not involved in dimer formation and are oriented towards the dimer's external environment, may evolve under positive selection, which would be consistent with an immune function.

To understand the evolutionary history of the molecule, I searched for Dscam related genes in representatives of chelicerates (*Ixodes scapularis*) and myriapodes (*Strigamia maritima*), two other groups of arthropods. In both myriapodes and chelicerates, Dscam diversified extensively by whole gene duplications and by duplications of some internal exons coding for one Ig domain region, but not several, like in insects and crustaceans. Similar duplications could have provided the raw material from which the highly diverse Dscam evolved uniquely in the ancestors of crustaceans and insects. I propose a speculative scenario under which the evolution of this remarkable gene might have occurred.

INTRODUCTION

Cell adhesion molecules were needed early in evolution for intercellular cohesion and communication of multicellular organisms (Hynes and Zhao 2000). Throughout the evolution of metazoans, cell adhesion molecules were recruited for many different cellular functions such as cell proliferation and differentiation, apoptosis, migration and parasite recognition (Buckley et al. 1998; Humphries and Newham 1998). Many members of this family are at least in part built from immunoglobulin domains (Ig) (Chothia and Jones 1997) and several show considerably high molecular diversity associated with alternative splicing (Kohmura et al. 1998; Wu and Maniatis 1999).

The Dscam gene

The Down syndrome cell adhesion molecule (Dscam) gene was first described in humans associated with defects in the nervous system (Yamakawa et al. 1998). Subsequently, several members of the Dscam family were describe in other metazoans, in which its main known function is related to the development of the nervous system (Schmucker et al. 2000; Agarwala et al. 2001; Fusaoka et al. 2006; whole Millard et al. 2007). Both vertebrates and insects have Dscam members that resulted from gene duplications like DSCAM and DSCAM-like in humans and DscamL1, DscamL3 and DscamL4 in insects.

These proteins are typically cell surface receptors composed of 9(Ig)-4(FN)-Ig-2(FN) (Shapiro, Love, and Colman 2007), where FN stands for fibronectin type III domain. The extracellular domains are usually followed by a transmembrane domain and a cytoplasmic tail. One member of this family, named Dscam in insects, is the most remarkable example known of protein diversification by duplication and alternative splicing (AS) (Schmucker et al. 2000). The gene encoding this member of the Dscam family, evolved dozens of internal exon tandem duplications differing in amino acid composition and arranged in three arrays in the Dscam locus. The three arrays of exons encode half of the second and third Ig domains and the complete Ig7. This is made possible by a refined mechanism of mutually exclusive AS that ensures that in the mature mRNA only one exon per array is present.

Function of Dscam diversity Most of Dscam's diversity has been shown to be essential for the correct development of the nervous system in flies, suggesting that the isoforms are not redundant functionally (Chen et al. 2006). Homophilic binding between identical isoforms has been shown *in vitro*, indicating a degree of binding specificity in which 95% of all isoforms will bind only to other identical isoforms (Wojtowicz et al. 2004; Wojtowicz et al. 2007). This homophilic binding allows *in vivo*, that nervous cells recognize each other leading to a self-avoidance behavior that is at the basis of neural wiring in *Drosophila melanogaster* (Hughes et al. 2007; Matthews et al. 2007; Soba et al. 2007).

The diversity of Dscam isoforms has been suggested furthermore to be involved in immunity of insects (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006). Knocking down Dscam by RNAi in third instar larvae of *Drosophila melanogaster* and in *Anopheles gambiae* immune competent Su5B cells, reduces phagocytosis by 45 to 60% (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006). *Anopheles* mosquitos depleted of Dscam through gene silencing, suffered from high microbe proliferation in the hemolymph even in the absence of experimental challenge (Dong, Taylor, and Dimopoulos 2006). Different Dscam isoforms have different binding affinities to bacteria (Watson et al. 2005) and in mosquito Su5B cells, isoforms induced by different pathogens had higher affinity for the inducer pathogen than for other pathogen species (Dong, Taylor, and Dimopoulos 2006). Contrastingly, another study has shown that null Dscam mutant *D. melanogaster* embryonic hemocytes were still able to phagocyte bacteria as efficiently as their wild counterparts (Vlisidou et al. 2009). A feature that is very suggestive of an immune role of Dscam, is the fact that soluble isoforms produced by the fat body of flies and mosquitos circulate in the hemolymph where they could mediate opsonization (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006).

Strutural aspects of Dscam The structure of the first eight Ig domains of Dscam has been elucidated. The first four Ig domains adopt a so called horse-shoe conformation (Meijers et al. 2007). The horseshoe conformation seems to create singular adhesive properties given that it is common to other cell adhesion molecules involved both in the nervous system like axonin, and in the immune system like hemolin (Su et al. 1998; Schurmann et al. 2001; Meijers et al. 2007). In hemolin this structure has been shown to create a binding site to bacterial lipopolysaccharides (Su et al. 1998). The remaining four Ig domains (Ig5 to Ig8) provide the molecule with a serpentine shape (S shape) (Sawaya et al. 2008). The homophilic binding between identical isoform occurs through the formation of Dscam dimers (Fig. 1).

Remarkably, the Dscam regions involved in dimer formation are segments of Ig2, Ig3 and Ig7 domains coded by the alternative exons (Meijers et al. 2007; Sawaya et al. 2008). In this way the genetic diversification caused by the duplications, coupled with the strong specificity of Dscam's homophilic binding, provide a highly diverse "key-lock" system which nervous cells exploit extensively (Hughes et al. 2007; Matthews et al. 2007; Meijers et al. 2007; Soba et al. 2007; Sawaya et al. 2008).

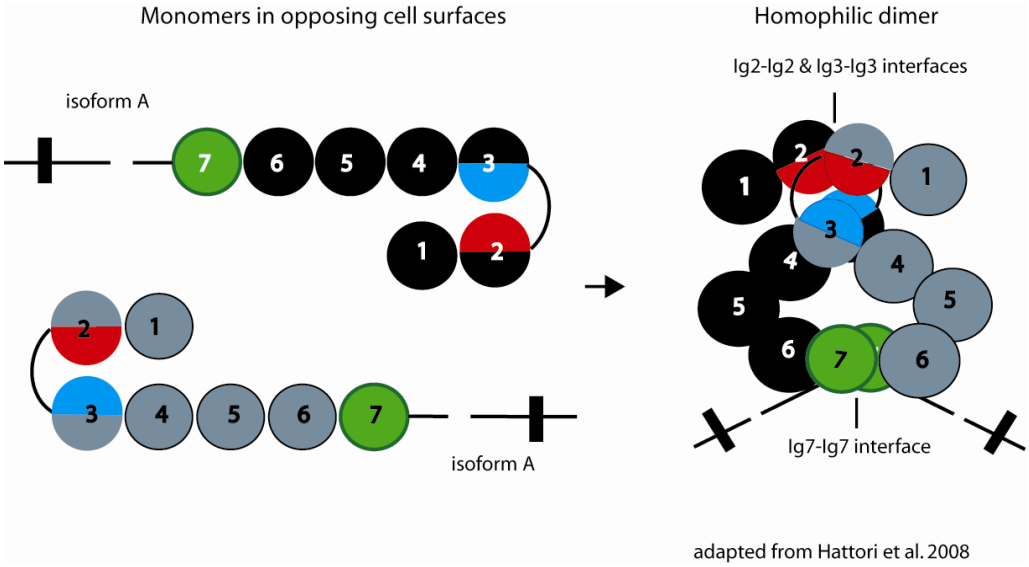


Figure 1 Model based on the $Dscam_{1\text{-}8}$ crystal structure for the conformation of the first seven Ig domains of Dscam in monomers (right) and after the formation of dimers (left). In monomers, the first four Ig domains form a compact horse-shoe structure whereas the remaining Ig domains have a flexible structure. Upon homophilic binding between identical isoforms (here, isoform A) mediated by the variable regions of Ig2, Ig3 and Ig7 (in color) the dimer acquires an S shape.

The implications of the structural features above described for an immune role of the molecule have not been tested. Nevertheless, it has been suggested that certain variable regions of Ig2 and Ig3 that are not involved in the formation of dimers, could recognize pathogen-associated molecular patterns (Meijers et al. 2007).

Dscam mutually exclusive alternative splicing Although the mechanisms of mutually exclusive alternative splicing of the duplicated exons are not fully understood, a few features within the Dscam gene have been identified in *Drosophila*. One feature is a secondary structure formed by the intron just preceding the first alternative exon coding for half of Ig2 (exon 4). This is a helical structure (iStem) that has been determined to be important in regulating the inclusion of exons 4 in the mRNA (Kreahling and Graveley 2005). Other features have been identified that regulate the array of exons 6 (Graveley 2005), namely two conserved sequence elements: the docking site and the selector site. The first is located in the intron between the constitutive exon 5 and the first exon 6 (which codes for half of Ig3 domain), and the second is located upstream of each alternative exon 6. Importantly, the selector sequence is complementary to the docking site sequence, and (Graveley 2005) suggested that the interaction between these two sites could be part of the mechanism ensuring that only one exon 6 is included in the mRNA, although this has not been demonstrated. The region of duplicated exons coding for the Ig7 domain has not been analyzed so far.

Dscam exon duplications The alternative exons have arisen by reiterative exon duplication and deletion in the three arrays. In the majority of cases, exons that are proximal within the array are more similar to each other than to the remaining exons. This has been suggested to result from frequent recombination between similar exons and to occur more frequently in the central regions than in the ends of the array (Graveley et al. 2004; Lee et al. 2009). Despite the similarities in the apparent mechanism of duplication, the three arrays seem to have undergone different patterns of exon radiation; exons 4 have duplicated notoriously less than the exons forming the other two arrays (Crayton et al. 2006; Lee et al. 2009).

This study

I aimed at elucidating the evolutionary history of the variable Dscam gene and at understanding how that relates to the different functions of the molecule. To pursue that, I have used sequence comparative analysis, quantification of Dscam expression, phylogenetic, molecular evolution and population genetics tools. Initially I started by studying Dscam in the closest relatives to insects, the brachiopod crustaceans (Glenner et al. 2006), using the species *Daphnia magna* and *Daphnia pulex*. I also used the species *Daphnia magna* for studying the expression of Dscam in

relation to parasitism. To approach questions related to the molecular evolution of regions of the gene involved in dimer formation and other regions putatively involved in parasite recognition, I have analyzed those regions in different populations of *Daphnia magna* and in several species of *Daphnia* and *Drosophila*. Finally, to trace the evolutionary history of the gene I did a comparison of several metazoan species, with a particular focus on the arthropod phylum by studying Dscam in representatives of chelicerates and myriapods.

REFERENCES

Agarwala, K. L., G. Subramaniam, Y. Tsutsumi, T. Suzuki, A. Kenji, and K. Yamakawa. 2001. Cloning und Functional Characterization of DSCAML1, a Novel DSCAM-like Cell Adhesion Molecule that Mediates Homophilic Intercellular Adhesion. Biochem Bioph Res Co:760-772.

Buckley, C. D., G. E. Rainger, P. F. Bradfield, G. B. Nash, and D. L. Simmons. 1998. Cell adhesion: more than just glue (Review). Molecular Membrane Biology 15:167-176.

Chen, B. E., M. Kondo, A. Garnier, F. L. Watson, R. Püettmann-Holgado, D. R. Lamar, and D. Schmucker. 2006. The Molecular Diversity of Dscam Is Functionally Required for Neuronal Wiring Specificity in Drosophila. Cell 125:607-620.

Chothia, C., and E. Y. Jones. 1997. The molecular structure of cell adhesion molecules. Annual Review of Biochemistry 66:823-862.

Crayton, M. E., 3rd, B. C. Powell, T. J. Vision, and M. C. Giddings. 2006. Tracking the evolution of alternatively spliced exons within the Dscam family. BMC Evol Biol 6:16.

Dong, Y., H. E. Taylor, and G. Dimopoulos. 2006. AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biol 4:e229-.

Fusaoka, E., T. Inoue, K. Mineta, K. Agata, and K. Takeuchi. 2006. Structure and function of primitive immunoglobulin superfamily neural cell adhesion molecules: a lesson from studies on planarian. Genes to Cells 11:541-555.

Glenner, H., P. F. Thomsen, M. B. Hebsgaard, M. V. Sorensen, and E. Willerslev. 2006. The origin of insects. Science 314:1883-1884.

Graveley, B., K. Amardeep, G. Dorian, Z. S. Lawrence, R. Lee, and C. J. c. 2004. The organization and evolution of the Dipteran and Hymenopteran Down syndrome cell adhesion molecule (*Dscam*) genes. RNA:1499:1506.

Graveley, B. R. 2005. Mutually exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell 123:65-73.

Hughes, M. E., R. Bortnick, A. Tsubouchi, P. Baumer, M. Kondo, T. Uemura, and D. Schmucker. 2007. Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron 54:417-427.

Humphries, M. J., and P. Newham. 1998. The structure of cell-adhesion molecules. Trends in Cell Biology 8:78-83.

Hynes, R. O., and Q. Zhao. 2000. The evolution of cell adhesion. Journal of Cell Biology 150:F89-F95.

Kohmura, N., K. Senzaki, S. Hamada, N. Kai, R. Yasuda, M. Watanabe, H. Ishii, M. Yasuda, M. Mishina, and T. Yagi. 1998. Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. Neuron 20:1137-1151.

Kreahling, J. M., and B. Graveley. 2005. The iStem, a Long- Range RNA Seconday Structure Element Required for Efficient Exon Inclusion in the *Drosophila Dscam* Pre-mRNA. Molecular and Celular Biology 25:10251-10260.

Lee, C., N. Kim, M. Roy, and B. R. Graveley. 2009. Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. Rna 16:91-105.

Matthews, B. J., M. E. Kim, J. J. Flanagan, D. Hattori, J. C. Clemens, S. L. Zipursky, and W. B. Grueber. 2007. Dendrite self-avoidance is controlled by Dscam. Cell 129:593-604.

Meijers, R., R. Puettmann-Holgado, G. Skiniotis, J.-h. Liu, T. Walz, J.-h. Wang, and D. Schmucker. 2007. Structural basis of Dscam isoform specificity. Nature 449:487-491.

Millard, S. S., J. J. Flanagan, K. S. Pappu, W. Wu, and S. L. Zipursky. 2007. Dscam2 mediates axonal tiling in the Drosophila visual system. Nature 447:720-U714.

Sawaya, M. R., W. M. Wojtowicz, I. Andre, B. Qian, W. Wu, D. Baker, D. Eisenberg, and S. L. Zipursky. 2008. A double S shape provides the structural basis for the extraordinary binding specificity of Dscam isoforms. Cell 134:1007-1018.

Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. l. Zypursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity Cell 101:671-684.

Schurmann, G., J. Haspel, M. Grumet, and H. P. Erickson. 2001. Cell adhesion molecule L1 in folded (Horseshoe) and extended conformations. Molecular Biology of the Cell 12:1765-1773.

Shapiro, L., J. Love, and D. R. Colman. 2007. Adhesion molecules in the nervous system: Structural insights into function and diversity. Annual Review of Neuroscience 30:451-474.

Soba, P., S. Zhu, K. Emoto, S. Younger, S. J. Yang, H. H. Yu, T. Lee, L. Y. Jan, and Y. N. Jan. 2007. Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. Neuron 54:403-416.

Su, X. D., L. N. Gastinel, D. E. Vaughn, I. Faye, P. Poon, and P. J. Bjorkman. 1998. Crystal structure of hemolin: A horseshoe shape with implications for homophilic adhesion. Science 281:991-995.

Vlisidou, I., A. J. Dowling, I. R. Evans, N. Waterfield, R. H. ffrench-Constant, and W. Wood. 2009. Drosophila embryos as model systems for monitoring bacterial infection in real time. PLoS Pathog 5:e1000518.

Watson, L. F., F. T. Püttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel, and D. Schmucker. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects Science 309:1874-1878

Wojtowicz, W. M., J. J. Flanagan, S. S. Millard, and S. L. Zipursky. 2004. Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. Cell 118:619-633.

Wojtowicz, W. M., W. Wu, I. Andre, B. Qian, D. Baker, and S. L. Zipursky. 2007. A vast repertoire of Dscam binding specificities arises from modular interactions of variable ig domains. Cell 130:1134-1145.

Wu, Q., and T. Maniatis. 1999. A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell 97:779-790.

Yamakawa, K., Y.-K. Huo, M. A. Haendel, R. Hubert, X.-N. Chen, G. E. Lyons, and J. R. Korenberg. 1998. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. Hum Mol Genet 7:227-237.

CHAPTER 1

THE DSCAM HOMOLOGUE OF THE CRUSTACEAN *DAPHNIA* IS DIVERSIFIED BY ALTERNATIVE SPLICING LIKE IN INSECTS

Daniela Brites[*], Seanna McTaggart[*], Krystalynne Morris, Jobriah Anderson, Kelley Thomas, Isabelle Colson, Thomas Fabbro, Tom J. Little, Dieter Ebert and Louis Du Pasquier (2008). *Molecular Biology and Evolution*.25 (7):1429-1439.

[*]these authors contributed equally to this work.

ABSTRACT In insects, the homologue of the Down syndrome cell adhesion molecule (Dscam) is a unique case of a single-locus gene whose expression has extensive somatic diversification in both the nervous and immune systems. How this situation evolved is best understood through comparative studies. We describe structural, expression and evolutionary aspects of a Dscam homolog in 2 species of the crustacean *Daphnia*. The Dscam of *Daphnia* generates up to 13,000 different transcripts by the alternative splicing of variable exons. This extends the taxonomic range of a highly diversified Dscam beyond the insects. Additionally, we have identified 4 alternative forms of the cytoplasmic tail that generate isoforms with or without inhibitory or activating immunoreceptor tyrosine-based motifs (ITIM-ITAM), something not previously reported in insect's Dscam. In *Daphnia*, we detected exon usage variability in both the brain and hemocytes (the effector cells of immunity), suggesting that Dscam plays a role in the nervous and immune systems of crustaceans, as it does in insects. Phylogenetic analysis shows a high degree of amino acid conservation between *Daphnia* and insects except in the alternative exons, which diverge greatly between these taxa. Our analysis shows that the variable exons diverged before the split of the two *Daphnia* species and is in agreement with the nearest-neighbour model for the evolution of the alternative exons. The genealogy of the Dscam gene family from vertebrates and invertebrates confirmed that the highly diversified form of the gene evolved from a non-diversified form before the split of insects and crustaceans.

INTRODUCTION

The Down syndrome cell adhesion molecule (Dscam) belongs to a family of cell-membrane molecules involved in the differentiation of the nervous system. As with some other members of the family (e.g. Axonin, Roundabout, NCAM, contactin, L1CAM), the extracellular region of Dscam is made of Immunoglobulin (Ig) and Fibronectin (FN) domains. Throughout the metazoa, the *bona fide* Dscam domain composition and physical arrangement remains identical, namely, 9(Ig)-4(FN)-(Ig)-2(FN) (Shapiro et al., 2007)

For mammals and insects whose genome sequences are available, additional Dscam gene copies may be found. For example, humans have two gene copies, Dscam and the paralogue Dscam-Like1 (Dscam-L1) (Yamakawa et al.1998; Agarwala et al. 2001). Insects also have Dscam and several Dscam paralogs that have been named Dscam-L (Schmucker et al. 2000; Millard et al. 2007). In humans, the Dscam gene can generate three different transcripts through cryptic splicing sites in the gene (Yamakawa et al.1998). In contrast, the *Drosophila* Dscam, but not Dscam-L, has the potential to generate over 38,000 different transcripts (Schmucker et al. 2000). This unprecedented repertoire of transcripts is due to four arrays of alternative exons that are spliced together in a mutually exclusive manner. The alternative exons encode the first half of the second and third Ig domains, the entire seventh Ig domain, and the transmembrane segment.

In insects, the many different isoforms of Dscam play an essential role in growth and the directed extension of axon branches (Schmucker et al. 2000; Chen et al. 2006; Hattori et al. 2007). Biochemical studies support a model in which each isoform preferentially binds to the same isoform on opposing cell surfaces, providing neurons with a homolog interaction recognition system (Wojtowicz et al. 2004). In *Drosophila*, the diversity of Dscam isoforms is necessary for neural wiring specificity (Chen et al. 2006; Hattori et al. 2007), but is also thought to be important in insect immunity. For example, Dscam transcripts are found in hemocytes, in cells from the fat body, a central organ involved in immunity, and soluble Dscam molecules are present in the hemolymph serum (Watson et al. 2005). Additionally, the silencing of Dscam by RNAi reduces the ability of *Drosophila* hemocytes to phagocytose by ~60% (Watson et al. 2005), while in mosquitoes it results in reduced survival after pathogen exposure (Dong, Taylor and Dimopoulos 2006). Watson et al (2005) demonstrated that Dscam binds to bacteria and that this capacity varies among isoforms (Watson et al. 2005). Finally, different splice variant repertoires are expressed between pathogen-challenged and unchallenged mosquitoes and cell lines (Dong, Taylor and Dimopoulos 2006).

A Dscam gene with alternative spliced exons generating three hypervariable Ig domains has evolved in several insect orders over ~250 million

years (Graveley et al. 2004; Watson et al. 2005). The origin of the alternative spliced exons remains elusive as, generally, no homology was found outside of insects (Crayton et al.2006). Here we describe a homolog of a diversified Dscam in the branchiopod Crustacean *Daphnia. Daphnia* reproduce mostly clonally, which permits us to study Dscam expression with strict control of the genetic background. The Dscam gene was studied in two different species, *Daphnia magna* and *Daphnia pulex*, which are thought to have diverged approximately 200 My ago (Colbourne and Hebert 1996). Recent studies suggest that hexapodes (arthropods having six legs, including insects) and branchiopod crustaceans are sister groups that shared a common ancestor around 420 My ago (Glenner et al. 2006). Thus, the description and phylogenetic comparison of the Dscam gene across insects and crustaceans can provide insight into the evolution of the gene and the origin of its dual function in the nervous and immune systems. Furthermore, closer examination of the patterns of sequence evolution of the alternative exons within and between species, provide insights into the evolution of the alternative exons.

MATERIAL AND METHODS

Gene recovery We used insect Dscam protein sequences to probe the *D. pulex arenata* (http://daphnia.cgb.indiana.edu/) scaffolding 10X using tBLASTn (Altschul et. al 1997). We extracted the region of scaffolding corresponding to significant matches, plus an additional 2000 nt up and downstream. This sequence was manually annotated in Artemis (http://www.sanger.ac.uk/Software/Artemis) using BLAST high scoring segment pairs from the initial tBLASTn search, in addition to those obtained from BLASTp searches of the open reading frames of the target scaffold sequence in all three frames of the translated sequence, %GC content, and the identification of GT-AG boundaries that frame introns. We used the annotated gene as a new query amino acid sequence to search the *Daphnia* genome assembly for any additional copies.

We accepted genes as Dscam paralogs if, according to the SMART database, their extracellular Dscam domain structure was 9(Ig)-4(FN)-(Ig)-2(FN). The genome of *D. pulex* contains two regions with homology to non-variable Dscam genes. One of these lacks two Ig domains, the transmembrane segment, the cytoplasmic tail, and the initiator methionine could not be identified. The second region lacks one Ig and one Fn domain. The NCBI database was searched for additional putative Dscam homologs and paralogs (species accession numbers provided in the supplementary material). In *Drosophila* four Dscam members have been reported (Millard et al. 2007): the canonical variable Dscam (aaf71926.1) and the putative paralogues cg31190 (Dscam-L1), cg32387 (Dscam-L2) and cg 33274.

Only Dscam-L2 has a canonical Dscam domain structure and two alternatively spliced exons coding for the Ig 7 domain of the molecule. The predicted structure of cg33274 lacks one Ig domain and thus was excluded from further analysis. The presence of the first FN domain of Dscam-L1 is ambiguous, however the length of the gene is compatible with a full Dscam gene. Therefore, we included Dscam-L1 and Dscam-L2 in the Dscam paralog analysis.

We also sequenced Dscam from another *Daphnia* species, *D. magna*. Dscam genomic sequences were obtained from a fosmid library (see supplementary material for details). Additional genomic and cDNA data were generated from a single clonal line (clone Mu11, originally isolated from a pond near Munich, Germany). Further Dscam cDNA was obtained from hemocytes of the genetic line HO2 (originally isolated from a pound in Hungary) that were infected with the pathogenic bacteria *Pasteuria ramosa* (Ebert et al. 1996).

RNA extraction and cDNA synthesis *Daphnia magna* and *D. pulex* mRNA extractions were carried out with Dynalbeads technology (Dynalbeads mRNA Direct^{tm} Micro kit) following the manufacturer's instructions. For whole-body mRNA preparation, mRNA was eluted in 6µl of 10mM Tris-HCl and used to synthesize cDNA directly or frozen at –80°C. To obtain mRNA from hemocytes, single individuals were immobilized in microtest plates (Terasaki microtiter plates, GREINER BIO-ONE) with a drop of 0.75% agar at 37°C. Hemolymph was withdrawn by capillary action, with twice-pulled microcapillary glass tubes (Harvard apparatus GC100TF-10) inserted into the heart chamber and brains were dissected. Both tissue types were immediately stored in RNA*later* (Ambion) solution.

To obtain the 5' region of Dscam mRNA, we used SMART technology (SMART^{tm} RACE cDNA Amplification Kit, CLONTECH) on mRNA samples extracted from whole *D. magna*. We used 3µl of eluted mRNA with two reverse primers (primer sequences available upon request) specific to the Ig1 and Ig4 exons of *D. magna*. The remainder of the cDNA sequences were synthesized in a 20 µl reverse transcription (RT) reaction consisting of 2 µl of SuperScript^{tm}III Reverse Transcriptase (Invitrogen) and 1 µl of oligo(dT) (50 µM), following the instructions of the manufacturer. In the RT reactions, either 3 µl of mRNA were used or, in the case of hemocyte and brain preparations, the whole mRNA samples were used directly to make solid-phase first strand cDNA libraries.

PCR, cloning and sequencing To obtain the full Dscam cDNA sequence from *D. magna*, oligonucleotide primer pairs were designed using the *D. pulex* sequence in regions with high amino acid conservation among *D. pulex* and several insect species. PCR was carried out using the BD Advantage^{tm} 2 PCR Kit on 1 µl of cDNA according to the manufacturer's directions. Several PCR reactions were required in order to complete the cDNA sequence (primer sequences and PCR

conditions available upon request). To obtain the cDNA sequence of Ig2, Ig3 and Ig7 variable domains, we PCR amplified the first strand cDNA libraries prepared with the mRNA isolated from hemocytes and brain. Fifteen µl of the total 20 µl RT reaction were washed twice in 1x PCR buffer. The beads were combined with the PCR master mix and the reactions were submitted to the following PCR conditions: 95°C for 1 minute, 2 cycles of: 57°C for 30 seconds, 72°C for 5 minutes and 94°C for 2 minutes. The beads were then removed from the reactions, and the PCR proceeded as above for 35 cycles, except that the 72°C step was changed to 90 seconds. The PCR products were gel purified (QIAquick Gel Extraction kit, Qiagen) prior to cloning.

Most of the PCR products were cloned in the pCR 2.1- TOPO vector (Invitrogen). Due to the large size of the PCR product from the 3' RACE, it was cloned into a pCR-XL-TOPO vector (Invitrogen). All cloned products were sequenced under Big Dye terminator conditions, using the M13 reverse and/or M13 forward primers. For the PCR products that contained variable exons, several colonies were sequenced.

To test whether the exons from arrays 4, 6, and 11 are randomly expressed, we compared the observed frequency of the sequenced exons to the expected frequency using the Pearson chi-square statistic. The expected frequency was set to be equal for all exons present in the gene sequence. Simulations with the same number of replicates confirmed that the probability of a Type I error was always very close to 5%.

Genealogy of Dscam We constructed an amino acid multiple sequence alignment of the Ig and Fn domains for selected organisms. We did not include the cytoplasmic tail sequence as it is too divergent to align with confidence. We then created a Bayesian inference phylogeny using MrBayes 3.1.2. We used the mixed model option to choose the amino acid substitution model from each data set, a gamma rate distribution estimated from our dataset, and a burn-in equal to 1/10 the number of generations; after the burn-in phase every 100^{th} tree was saved. Two parallel Markov chains were run simultaneously in each of two runs. Tree length, amino acid model, log-likelihood score and alpha value of the gamma distribution were examined in the program Tracer v1.3 prior to the termination of MrBayes to ensure that all parameters had reached stationarity. All variable exons from each exon array were extracted from the genome sequence and aligned using the default parameters of the Clustalw program in MacVector (v7.2.3), where they were corrected by eye. Bayesian genealogies of each of the three variable exon arrays were constructed as described above for *D. magna* , *D. pulex* and *Apis melifera*.

To examine sequence divergence among exons within each array within and between the two *Daphnia* species, we computed the number of synonymous and nonsynonymous differences per synonymous (*ps*) and nonsynonymous site (*pn*) respectively. The calculations were performed using the Nei-Gojobori method

(Zhang, Rosenbergdagger and Nei 1998) estimating in all cases the transition/transversion ratio, using the pairwise deletion option and calculating standard errors by the bootstrap method (1000 replicates). These analyses were performed using the software MEGA version 4 (Tamura et al. 2007).

Nomenclature The major difference between Dscam family members is the presence or absence of arrays of alternatively spliced exons. For clarity, we shall refer to the gene with the alternative exon arrays as hypervariable Dscam and name it Dscam-hv.

RESULTS & DISCUSSION

Daphnia Dscam gene organization

The *Daphnia* Dscam-hv gene has a similar organization to its homolog in insects in that the exons coding for half of Ig domains 2 and 3 and the entire Ig 7 of the Dscam-hv protein are present in arrays of multiple exons (Fig. 1). The gene organization in both *Daphnia* species is very similar (*accession numbers: D. magna* EU307883, *D. pulex* EU307884). There are 82 exons present in *D. pulex* and 81 in *D. magna*, of which 32 exons account for the mature mRNA in both species (Fig. 1). They are organized as follows: the exon 4 array has 8 variants in both *Daphnia* species, the exon 6 array has 26 variants in *D. pulex* and 24 in *D. magna*, and the exon 11 array has 16 and 17 variants in *D. pulex* and *D. magna*, respectively (Fig.1). There are two main differences in the Dscam-hv gene arrangement between insects and *Daphnia*. First, insects have two alternatively spliced exon variants coding for the transmembrane domains, whereas *Daphnia* has only one (Fig. 1). Secondly, expression data revealed that 4 different cytoplasmic tails are expressed by both *Daphnia* species (Fig. 2A & B), whereas, to date, insects express only one cytoplasmic tail isoform. The cytoplasmic tail of *Daphnia* can be coded either by exons 26 to 31, or exon 30 can be skipped, which results in exon 31 being translated in a different reading frame (Fig. 2A). Furthermore, exon 27 may also be skipped accounting for two additional cytoplasmic tail possibilities. Altogether, the combined usage of the different alternatively spliced exons and cytoplasmic tail possibilities can potentially generate 13,312 different protein isomorphs in *D. pulex* and 13,056 in *D. magna*. This is the first finding of a Dscam-hv gene outside of the insects, and the first identification of alternative cytoplasmic tails in Dscam-hv.

Ig, Fn and the cytoplasmic tail domains of the Dscam protein

Dscam-hv amino acid sequence conservation is high between insects and *Daphnia* for most of the Ig and Fn domains, except for the regions

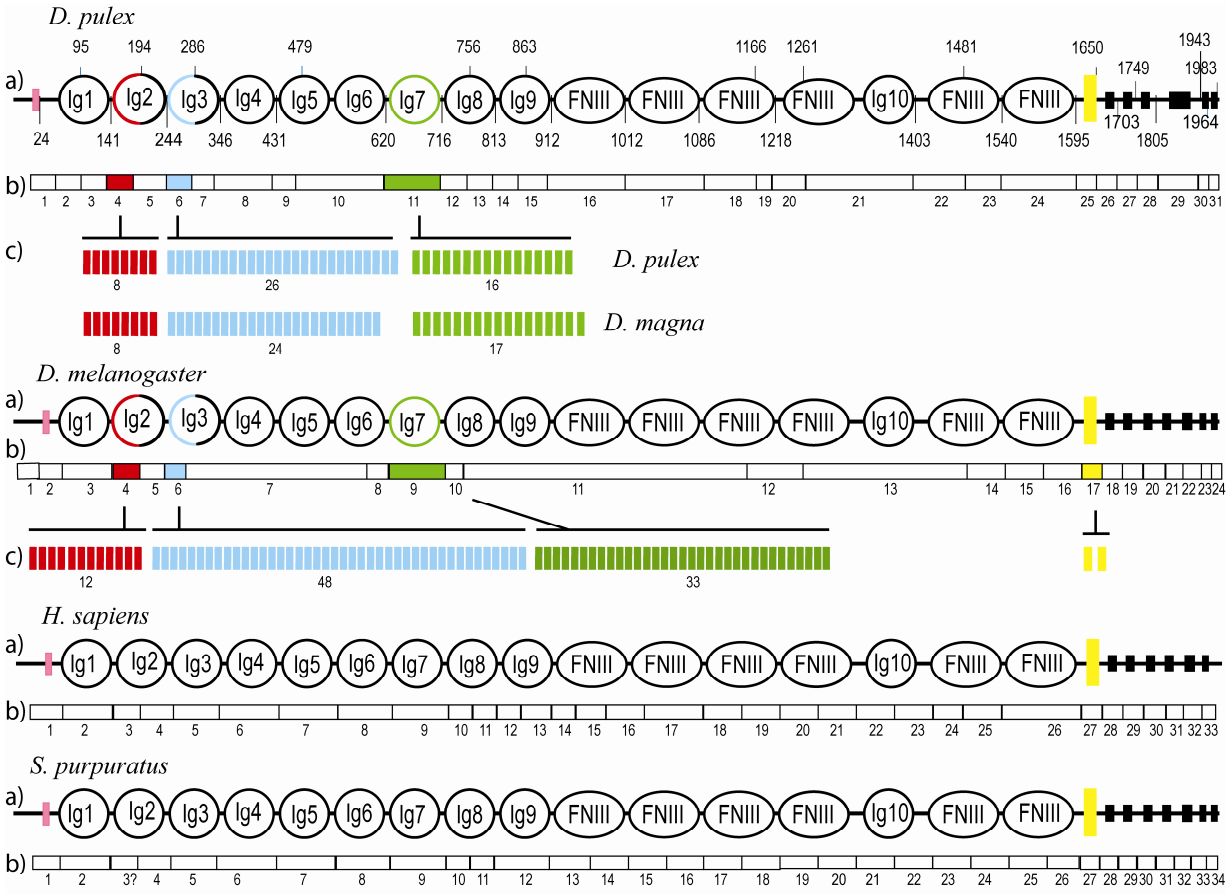


Figure 1 Dscam structure in *Daphnia*, *D. melanogaster*, *H.sapiens* and the sea urchin *Strongylocentrotus purpuratus*. a) protein domains, in *Daphnia* exon boundaries in the mRNA are indicated by amino acid numbers b) mRNA structure c) arrays of exons coding for the N- terminal parts of Ig2 (red) and Ig3 (blue) and the complete Ig7 (green) domains in *Drosophila* and *Daphnia* represented by bars that correspond to the number of alternative exons present in each species. The transmembrane domain (yellow) in *D. melanogaster* is coded by two alternative exons. The cDNA structure of *Strongylocentrotus purpuratus* between exon 2 and exon 4 is currently unclear.

coded by the alternative exons. Additionally, some highly conserved motifs are present in the cytoplasmic region of Dscam-hv in *Daphnia* and insects (Fig. 3), which are absent from Dscam or Dscam-L in insects. Schmucker et al. (2000) identified some of these conserved motifs as SH2/SH3 binding domains, which are involved in the binding of Pak to Dscam-hv *via* the adaptor protein Dock, that could mediate changes in the cytoskeleton of cells to promote axon guidance. While the strong similarity of these and other domains between *Daphnia* and insects (Fig. 3) indicates that the molecules interacting with Dscam-hv are likely the same in the two groups, the different cytoplasmic tails expressed by *Daphnia* show that differences also exist. Although the functional role of the different cytoplasmic tails is as yet unknown,

they are all expressed in both brain tissue and hemocytes. The 47 amino acids that may or may not be present in the cytoplasmic tail of *Daphnia*, depending on whether exon 27 is skipped, contain several short regions that are highly conserved between *Daphnia* and insects, namely an endocytosis/phagocytosis motif (YXXL, Fig. 3).

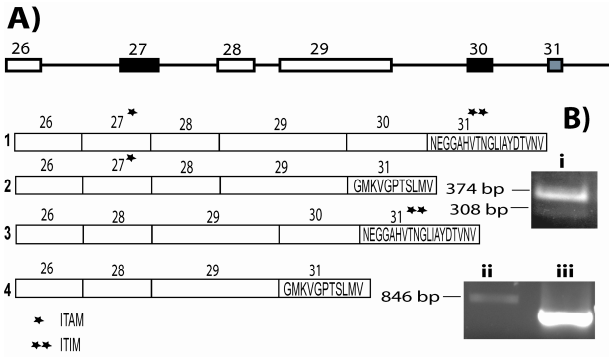


Figure 2 Schematic representation of *Daphnia* Dscam cytoplasmic tails A) *Daphnia magna* tail structure and splicing possibilities result in 4 alternative forms. Exons 26 to 31 code for the cytoplasmic tail. Exons 27 and 30 can be included in the mRNA or skipped. C-terminal end of the cytoplasmic tail changes if exon 30 is included (1), or skipped (3). Two other forms, (2) and (4), are obtained through the inclusion or exclusion of exon 27 B) *Daphnia magna* Dscam cytoplasmic tail expression in the whole body messenger RNA. i) The two bands correspond to the cDNA fragments that can be coded by exon 29 to exon 31. The bigger fragment includes exons 29, 30 and 31 and the smaller includes exons 29 and 31. ii) Fragment correspondent to cDNA containing exon 27 to exon 31. Cloning and sequencing of this fragment revealed that exon 30 may or may not be transcribed. iii) Control: whole body mRNA actin expression

In the two *Daphnia* species, this motif is part of a canonical ITAM, an immunoreceptor tyrosine-based activation motif (consensus: YXXL/V- 6 to 17 X- YXXL/V) (Barrow and Trowsdale 2006) (Fig. 3). Isoforms with or without these motifs may have very important differences in their signalling capacity and in regulating the expression of surface membrane receptors (Indik et al. 1995). The cytoplasmic tail variants that result from the inclusion or exclusion of exon 30 and the subsequent reading of exon 31 in two different reading frames, differ in length and in the composition of the PDZ (Postsynaptic density, disc large and zo-I protein domains) motif (Fanning and Anderson 1999; Sheng and Sala 2001) that occurs at the very end of the carboxyl end of each form. The alternative PDZ domains (YDTV if exon 30 is included, and SLMV if exon 30 is excluded (Fig. 2)) preferentially associate with different proteins and/or where they localize in the cellular membrane (Fanning and Anderson 1999). The longest form of the cytoplasmic tail of *D. magna* and *D. pulex* harbours an immune tyrosine-based inhibition motif (ITIM) (consensus: I/S/V/LXYXXV/L) (Fig. 2 and 3). After the interaction of the ligand with the extracellular part of the receptor, ITIM becomes phosphorylated on the tyrosine by Src kinases, which then allows it to recruit phosphotyrosine phosphatase that in turn decreases the activity of the cell (Barrow and Trowsdale 2006). The role of ITIM has not been investigated in any Dscam-hv, although the motif has been reported in mammalian Dscam (Staub, Rosenthal, and Hinzmann 2004). The fact that the alternative cytoplasmic tails in *Daphnia* may or may not encode an ITIM and ITAM (Fig. 2) suggests that they have very different signalling capacities.

Daphnia Dscam is therefore diverse in its recognition and effector capacities. The duality ITIM/ITAM in *Daphnia* Dscam reminds us of that observed in paired Ig receptors of vertebrates (Lanier 2001).

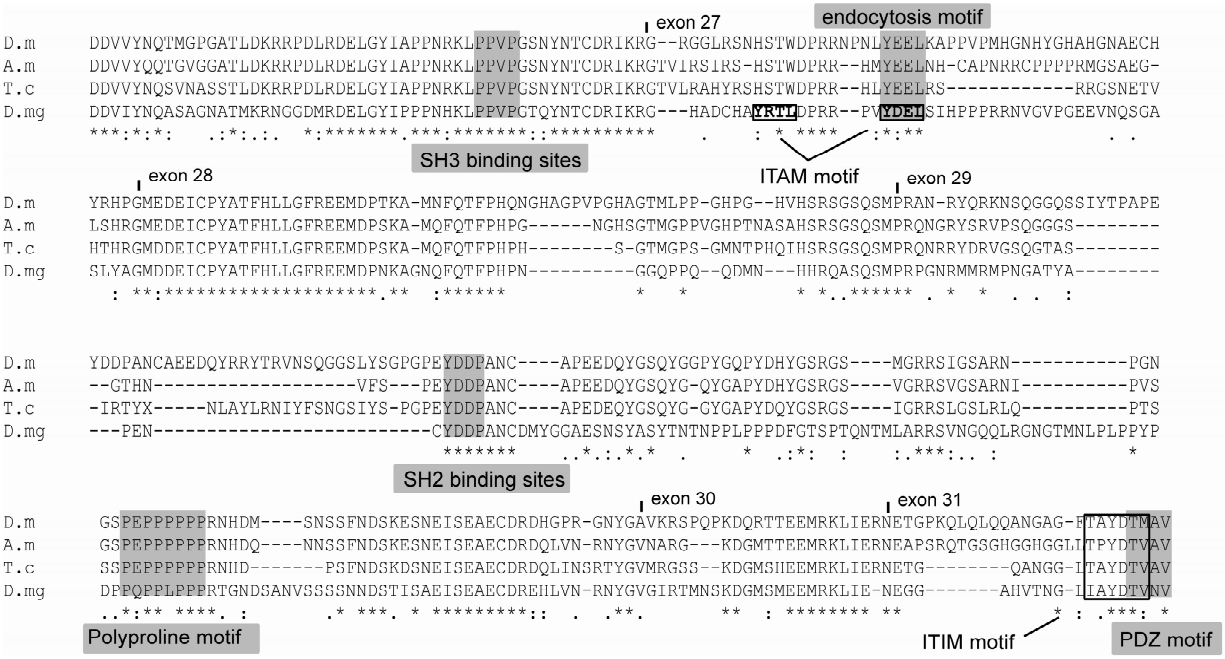


Figure 4 A) *Daphnia magna* expression of a Dscam region encompassing Ig3 to Ig7 in the brain and hemocytes. Sequencing revealed that each band is composed of many different isoforms corresponding to the expression of exon variants from arrays 4, 6 and 11. B) Exon usage frequency in different tissues in *D. magna*. Bars correspond to the expression of each exon in each tissue, relative to the total number of times the exon was observed in all tissues. C) Association of exons from each array in single mRNA molecules from brain, embryos and hemocytes. The bars on the right side of the graph represent the absolute number of times that each association was observed. Number of sequences: brain n=39; embryo n=16; hemocytes n=37. Exon 6.3 cannot be used because there is a mutation at the 3' end of the exon that does not allow splicing with exon 7 (splicing law changed from type 2 to type 0).

Expression of Dscam transcript diversity

To investigate how the potential exon diversity repertoire is expressed, we extracted mRNA from *D. magna* hemocytes, brain and whole embryos, using 10, 2, and 5 pooled *D. magna* individuals of the same clone respectively. From each of these extractions, we amplified, cloned and sequenced several RT-PCR products encompassing the three variable exon arrays. Variable expression of exons 4, 6 and 11 was detected in the hemocytes, brain and embryos (Fig. 4). All exons in the genomic sequence were expressed, except exons 6.3 and 6.10, demonstrating that *Daphnia* uses the full range of Dscam-hv diversity. The fact that various Dscam-hv isoforms are detected in both brain and hemocytes indicates that the Dscam-hv product diversity is exploited by both the

nervous and immune systems of *Daphnia*, as it is in insects.

Unlike *Drosophila*, which shows a more restricted expression of their exon 9 array (the equivalent to the exon 11 array in *Daphnia*), *Daphnia* has a restricted exon 6 array profile. Furthermore, more variants are expressed in brain tissue than in the hemocytes (Fig. 4). The restricted exon expression observed in *Daphnia* hemocytes could stem from the fact that the individuals examined were infected with one parasite, however, this result is consistent with those obtained from uninfected *Drosophila* (Watson et al. 2005). If each hemocyte expresses on average 14 different Dscam-hv isoforms, as in *Drosophila* (Neves et al. 2004), the restricted expression in hemocytes results in individual isoforms being present at a higher concentration, which may increase their functional capacity. Additionally, Dscam expression in hemocytes can be rapidly modulated following exposure to diverse pathogens (Dong, Taylor and Dimopoulos 2006), which implies a rapid turnover of expressed molecules. The numerous destabilizing RNA motifs (Bevilacqua, Ceriani and Capaccioli 2003) encountered in the 3'UTR of the *Daphnia* Dscam-hv could be related to this rapid turnover of the molecule (*D. magna*: 3 copies of ATTTA, 8 copies of TATT and 10 copies of TAAA in 1200 bp of 3'UTR; *D. pulex*: 6 copies of ATTTA, 20 copies of TATT, and 15 copies of TAAA within 2545 bp of the 3'UTR).

The observed expression patterns of exon arrays 4 and 11 in the brain do not significantly deviate from random expectation ($p=0.19$, $p=0.74$), but the expression pattern for exon 6 array does ($p=0.026$). In contrast, the expression pattern of exon arrays 4, 6 and 11 in hemocytes deviate strongly from random expectation ($p<0.0001$, $p=0.002$, $p<0.0001$). In both brain and hemocytes, the observed combinations of the three variable exons from one mRNA molecule deviate strongly from a random expectation ($p<0.0001$). Consistent with the hypothesis that the expression of Dscam-hv alternative exons is regulated, different exon combinations are preferred in the brain compared to hemocytes (Fig. 4). Previously, changes in Dscam-hv expression patterns for each exon across time, tissue and type of pathogen challenge have been demonstrated in both cell lines and in individuals of *Drosophila* and *Anopheles* (Celoto and Graveley 2001; Neves et al. 2004; Watson et al. 2005). Further immunological experiments will determine if this is also the case with *Daphnia*. Although the mechanisms for mutually exclusive splicing of the variable exons are not fully understood, studies of *Drosophila* have identified two sequence motifs within the Dscam-hv gene that appear to be involved in regulating exons from arrays 4 and 6 (Graveley 2005; Kreahling and Graveley 2005). These sequence motifs are also present in *Daphnia* (Fig. S1, Supplementary material), suggesting that the regulatory machinery is evolutionarily conserved between these taxa.

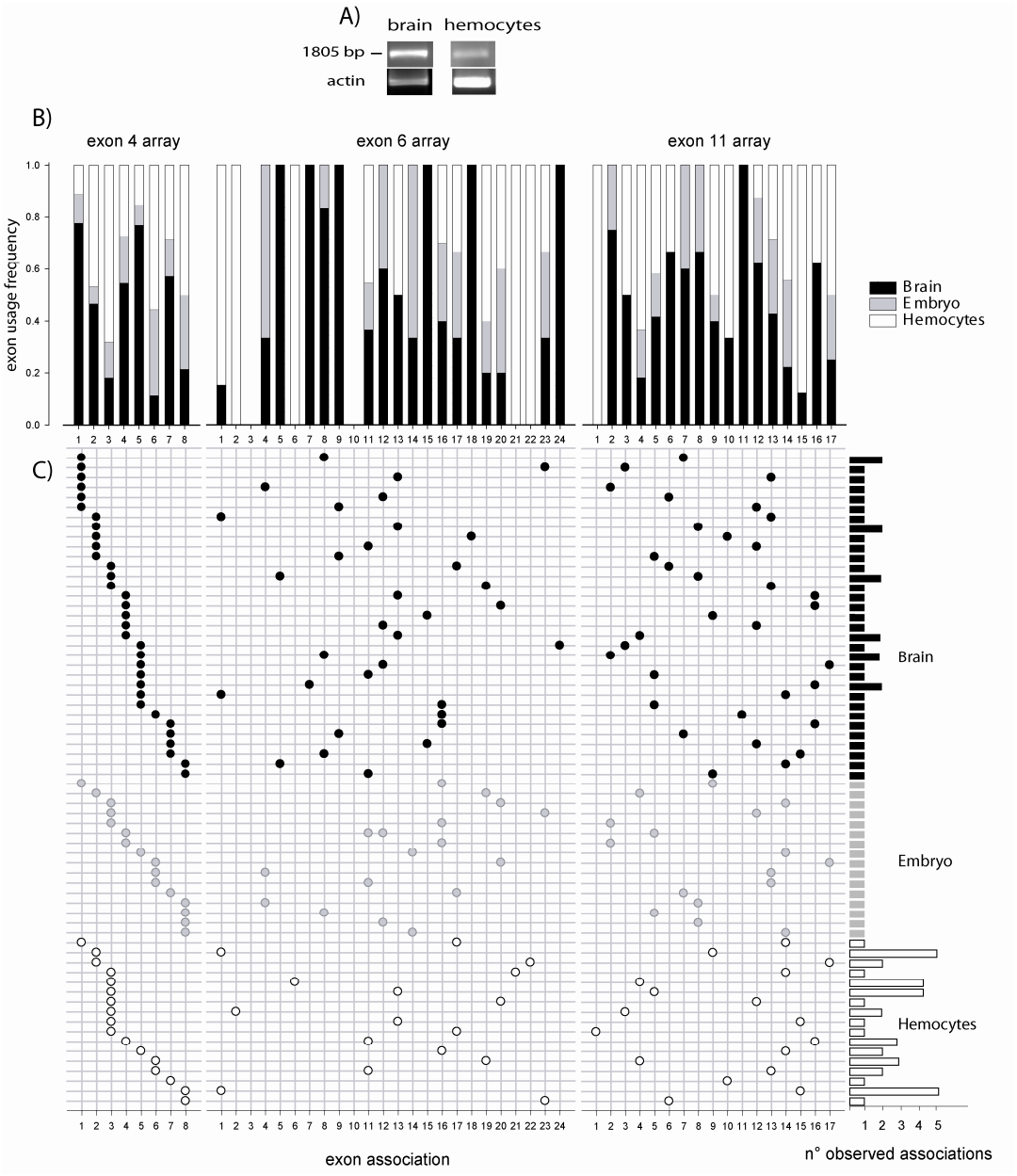


Figure 4 A) *Daphnia magna* expression of a Dscam region encompassing Ig3 to Ig7 in the brain and hemocytes. Sequencing revealed that each band is composed of many different isoforms corresponding to the expression of exon variants from arrays 4, 6 and 11. B) Exon usage frequency in different tissues in *D. magna*. Bars correspond to the expression of each exon in each tissue, relative to the total number of times the exon was observed in all tissues. C) Association of exons from each array in single mRNA molecules from brain, embryos and hemocytes. The bars on the right side of the graph represent the absolute number of times that each association was observed. Number of sequences: brain n=39; embryo n=16; hemocytes n=37. Exon 6.3 cannot be used because there is a mutation at the 3' end of the exon that does not allow splicing with exon 7 (splicing law changed from type 2 to type 0).

Variable regions within the alternative exons

A structural analysis of the first 4 Ig domains of two distinct Dscam-hv isoforms in *Drosophila* has demonstrated that the 5' portions of the alternative exons 4 and 6 contribute to regions of the protein that are essential for Dscam-hv homophilic binding and reside on a region called epitope I (Meijers et al. 2007). Located on the opposite side of the 3D structure of the molecule is epitope II, defined by the 3' region of exons 4 and the central region of exons 6. It does not participate in Dscam-hv homophilic binding (Meijers et al. 2007). A comparison of orthologous exons from arrays 4 and 6 from 12 *Drosophila* species revealed that the epitope II sequences are more variable than those of epitope I, suggesting that this region of the protein is under fewer selective constraints. Closer examination of the same sequences between *D. magna* and *D. pulex* is entirely consistent with the *Drosophila* observation, given that the regions of variability in crustaceans and insects are superimposable (Fig. S2, Supplementary material).

Phylogenies of the variable exons

Clear orthologs exist between the two *Daphnia* species for the vast majority of exons in each of the arrays (Fig. 5 A), meaning that interspecific sequence similarity is higher than intraspecific. This suggests that the occurrence of concerted evolution is not affecting the evolution of the multiple exons of each array in a significant way (Nei and Rooney 2005). This relationship is strongest in exon 4 array, where 1:1 orthologous pairs were identified for every exon (Fig. 5B). Similarly, almost all exon 6 array members have a clear pairing between the two *Daphnia* species (Fig. 5B), despite having different numbers of exons. These results are consistent with those obtained among three species of *Drosophila* (Graveley 2004). Sites of recent gene duplication of exon 6 variants in *D. pulex*, or gene loss in *D. magna*, are exons 12, 13 or 14 and exon 23 according to the numbering of *D. pulex* (Fig. 5B). Variation in exon 6 copy number also exists between *D. melanogaster* and *D. virilis* (48 and 52 copies respectively), indicating that recombination leading to exon loss/gain in this portion of the gene may be more frequent than in the exon 4 region. Regarding the exon 11 array, there have been two exon duplication/loss events since the split between the *D. pulex* and *D. magna* (Fig. 5B). In one case, *D. pulex* exon 11.5 does not have an orthologous match in *D. magna*. Since 1:1 orthologous pairings between the two Daphniids continue downstream, it is more likely that the *D. pulex* exon 11.5 is the result of an exon duplication event, as opposed to exon loss, in *D. magna*. In the other case, *D. magna* exons 11.13 and 11.14 are more closely related to each other than to any *D. pulex* exon, and thus likely arose by exon duplication in *D. magna* after the split between these two species. The fact that,

generally, orthology of the alternative exons has been maintained between the two *Daphnia* species, coupled with their short branch lengths, suggests that at least part of the exon sequence variation may be functionally contrained.

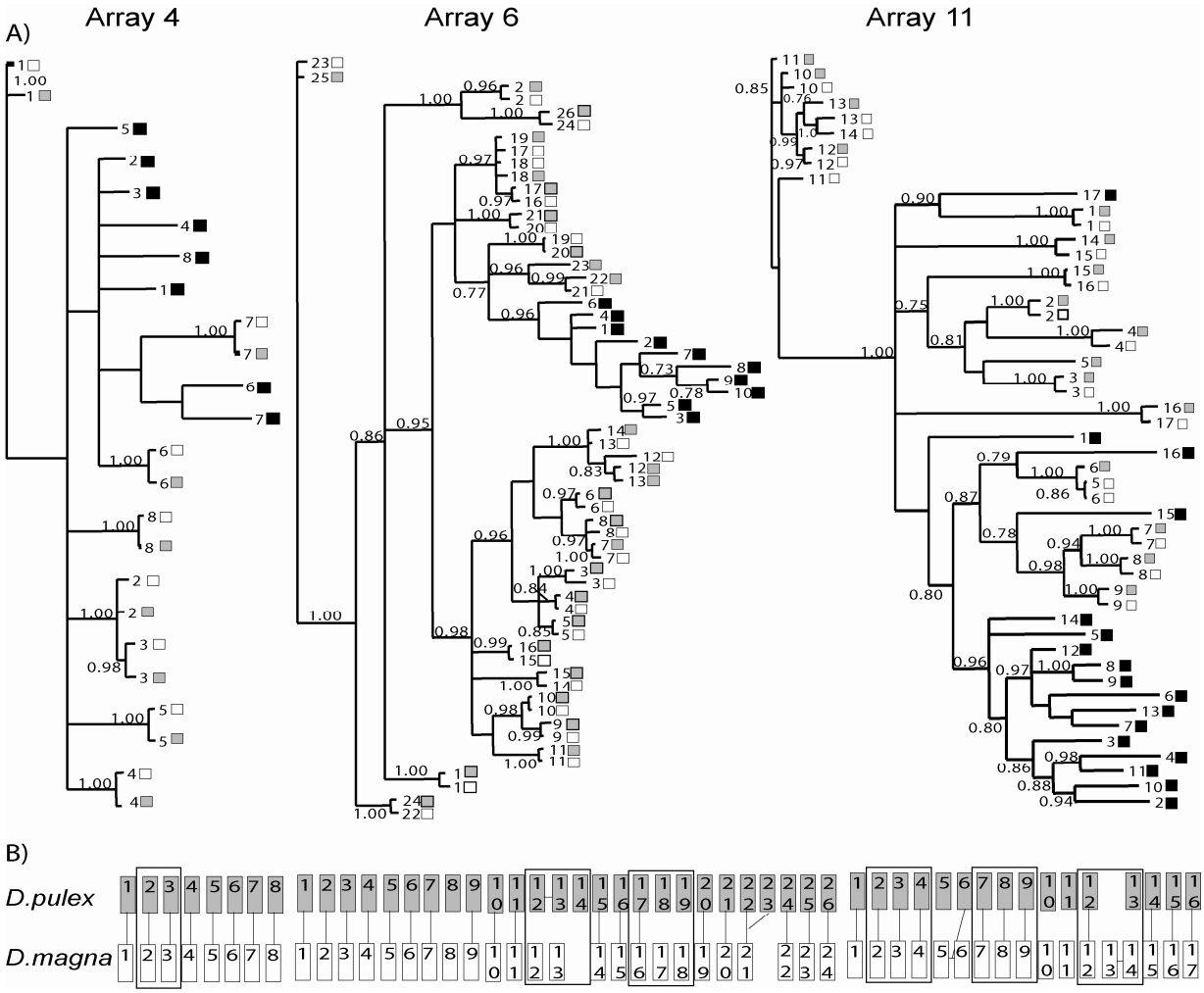


Figure 5 A) Bayesian analysis of the exons from *Daphnia magna* (white), *Daphnia pulex* (gray) and *Apis mellifera* (black) contained in the three variable arrays of the *Daphnia* Dscam gene. In the exon 6 tree, only 10 representatives of *A. mellifera* were included. B) Schematic representation of the exons depicting the orthologous pairing and synteny of the variable exons between the two *Daphnia* species. Boxes represent clustering among the nearest neighbors with a probability of 0.9 or more.

On the other hand, based on the lack of orthology between the alternative exons of *Daphnia* and insects (represented by *A. mellifera*, the insect species with the highest Dscam sequence similarity to *Daphnia*) (Fig. 5A), this constraint appears to be taxon specific. This contrasts with the high degree of sequence conservation in the constant domains of the molecule between these two groups of Arthropods. Furthermore, some characteristics of

each of the three arrays are consistently shared among species. For example, the exon 4 array always has fewer variants than either of the other two arrays. Such shared characteristics among the arrays could reflect that they have experienced similar selective constraints in both insects and crustaceans.

The evolution of the duplicated exons

It has been proposed that the alternative exons originated by duplication in a nearest-neighbour scenario, where exons closer to one another along the chromosome are more similar than exons that are further apart (Graveley et al. 2004). The phylogenies of the variable exon arrays 6 and 11 of the two *Daphnia* species are generally consistent with this model (Fig. 5). For example, in the exon 6 array some resolution beyond the orthologous pairings is obtained, where at least one large clade containing all the central exons in the array is strongly supported. Within this central exon clade, there are two additional clades that cluster exons 6.3-6.16 and 6.17-6.23 (numbering according to *D. pulex*) (Fig. 5A). The resolved members within the exon 11 array also correspond with the nearest neighbour hypothesis. However, in contrast, the exons present at the end and at the beginning of array 6 are more dissimilar to the central cluster. Furthermore, the relationship among paralogous exons is not well resolved for array 4, where only exon pairs 4.2 and 4.3 cluster together (Fig. 5A), suggesting that the exons in this cluster evolved rapidly, or that this array is older than the other two.

The number of synonymous substitutions per synonymous sites (*ps*) and nonsynonymous substitutions per nonsynonymous sites (*pn*) between alternative exons within each array is higher between than within the two *Daphnia* species (Fig. 6 and Fig. S3).

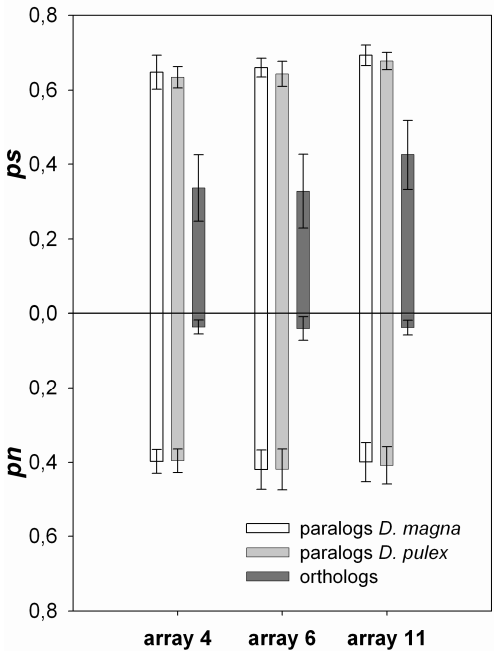


Figure 6 Average *ps* and *pn* of paralogs and orthologs from arrays 4, 6 and 11. The error bars correspond to the standard deviation of paralog and ortholog *ps* and *pn* values. The matrices of *ps* and *pn* values of all pairs of paralogs and orthologs and the estimated standard error are available by request.

This suggests that paralogs largely evolved according to the birth-and-death model, which assumes that new genes are created by repeated duplication events and that some duplicates may stay in the genome for a long time, whereas others are deleted or become non-functional (Nei

and Hughes 1992; Nei, Rogozin, and Piontkivska 2000). The recent exon duplication and deletions described for arrays 6 and 11 give further support to the appropriateness of this model in explaining how the variable Dscam arrays are evolving. Only one non-functional exon was found (see legend Fig. 5). The *ps* values between paralogs in one array are generally near the saturation level with most values between 0.4 and 0.7, whereas *ps* of orthologs although high, are lower (0.2-0.4) (See Fig. 6 for average values and Fig. S3). The number of nonsynonymous differences between paralogous and orthologous exons indicates that there are many more nonsynonymous differences between paralogs (*pn*: 0.1 to 0.6) than orthologs (*pn*: 0 to 0.06) and this pattern is very consistent in the three arrays (Fig. 6 for average values and Fig. S3). This difference in the number of substitutions in orthologs and paralogs for the three arrays supports that the duplicated exons in each cluster had already diverged in the ancestor of the two *Daphnia* species. The *dn* and *ds* values were calculated for orthologous exons by correcting the *ps* and *pn* values with the Jukes-Kantor formula (Ota and Nei 1994). The dn/ds ratio of orthologous exons indicates that strong selection is acting to maintain the amino acid composition of each exon (average dn/ds: array 4=0.08; array 6=0.1; array 11=0.06), Table S1). Selection acting upon paralogs in each array seems to have been much weaker, allowing for more nonsynonymous substitutions (Fig. 6) and subsequent diversification.

Dscam family evolution

Our searches for Dscam genes confirmed that, to date, only members of the insects (Crayton et al. 2006) and *Daphnia* have a Dscam-hv gene that contains at least three arrays of alternative exons (Fig. 1 & Fig. 7). We found no *sensu stricto* Dscam-L paralogs in the current *D. pulex* genome assembly, even though two genes with homology were found with a different domain organization (see material and methods section). Our tree shows that the vertebrate Dscam and Dscam-L genes are clearly separate from those of insects, the sea urchin and the flatworm *Dugesia*, despite the fact that the Dscam-L exon structure of insects lacks variable exon arrays, and thus superficially more closely resembles the vertebrate homologs (Fig. 7). Therefore, it seems that the ancestral Dscam gene duplicated in the two groups independently of one another, or that concerted evolution within the two groups has destroyed the phylogenetic signal at this deep level. The intron/exon boundaries of both vertebrate and insect Dscam gene copies also support the hypothesis of independent duplication, with insect Dscam-L genes intron/exon boundaries being more similar to those of Dscam-hv than to human Dscam or Dscam-L. Furthermore, the motifs identified by Crayton et al. (2006) that

discriminate the Dscam and Dscam-L of vertebrates were not found in any of the invertebrate Dscam genes. With respect to the timing of the duplication event within the invertebrates, both crustaceans and insects share the complex trait of alternative exon arrays, and likely the same mechanisms of mutually exclusive splicing, suggesting that the duplication event in the invertebrate lineage must have occurred before the split of the Pancrustaceans (Fig. 7). *Daphnia* appear to have strongly modified or lost its paralog of Dscam-hv. The two nematode genome sequences currently available (*C. elegans* and *C. briggsiae*) and the tunicate *Ciona* (a deuterostome) appear to lack Dscam altogether.

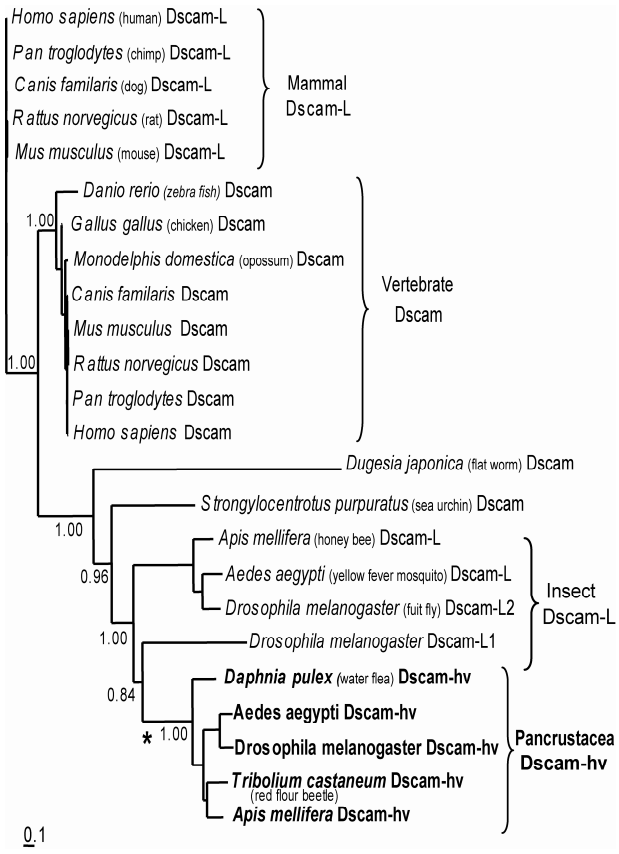


Figure 7 Bayesian topology of the extracellular regions of Dscam and Dscam–L genes from representative metazoan. Numbers at nodes are posterior probabilities. Only nodes relevant to the discussion are labeled. * represents the possible origin of mutually alternative splicing in Dscam.

Differences between the Dscam-hv, Dscam and Dscam-L can also been seen at the predicted properties of the respective proteins coded by these genes, like the number of gylocosylation sites. Glycosylation patterns suggest that there are fewer glycosylation sites in Dscam-hv compared to Dscam or Dscam-L (Table S2). This pattern holds true for the three insect species for which both forms of the gene occur, and for which sequences are available. Carbohydrates mediate interactions between recognition molecules and a great variety of glycan chains, and play a role in both the nervous and immune systems (Kleene and Schachner 2004). The higher number of glycosylation sites of the non-variable and Dscam-L proteins might be a functional alternative or complement the Dscam-hv molecules diversified by mutually alternative splicing.

CONCLUDING REMARKS

Alternative exons coding for Dscam-hv Ig domains are present in insects and in the crustacean *Daphnia*, but not in other invertebrates or vertebrates, suggesting that it evolved in the ancestor of the pancrustaceans. Dscam-hv amino acid conservation is high among divergent taxa, except in the regions that are coded by the alternative exons, which vary considerably in number and sequence between *Daphnia* and insects, and even among insects. Another level of variability in the alternative exons is evident when comparing more closely related species in the regions of Dscam-hv suspected to play a role in heterologous recognition (Meijers et al. 2007).

The structural position where this variability occurs seems to be conserved between *Daphnia* and several *Drosophila* species, despite the sequence divergence of their alternative exons. Thus, the principles underlying Dscam-hv diversity are conserved between *Daphnia* and insects. Furthermore, as in insects, *Daphnia* expresses diverse repertoires of Dscam-hv isoforms in both brain tissue and hemocytes. It is not known whether Dscam-hv diversity originally evolved by selection on the nervous system, the immune system, or both (Du Pasquier 2005).

Two non-exclusive selective advantages may be conferred to both the nervous and immune systems as a result of Dscam-hv diversity. First, it is beneficial to have a large number of different isoforms present in either system, even if their sole property is that they undergo homologous binding. This benefit has been demonstrated in the nervous system (Chen et al. 2006; Hattori et al. 2007), where the structural basis for homologous interactions is understood (Meijers et al. 2007). Specifically, the homologous interactions and their variegated expression on the cell surface allow large numbers of cells to be distinguished from one another. Similarly, the immune system could benefit by creating individualized hemocytes that can patrol without aggregating. If this is the case, many exons with different sequences, but not the precise exon sequences, would confer a selective advantage.

A second hypothesis is that isoforms are selected for their ability to bind to heterologous ligands, e.g. pathogens. In this scenario, specific exon sequences would be selected. Soluble forms of Dscam-hv circulate in the hemolymph of insects where they are unlikely to play any role in the nervous system, but could act as opsonins. Supporting this idea, inhibition of their expression results in a lower phagocytosis capacity and Dscam-hv isoform expression changes after exposure to various antigens (Dong, Taylor, and Dimopoulos 2006). Furthermore, a variable site on the molecule is oriented in a way that permits heterologous interaction (Meijers et al. 2007). All this suggests that the variability of Dscam-hv may be useful or even essential to the immune system.

In fact, the pattern of rapid evolution of the alternative exons in different species is reminiscent of Igsf members involved in innate immunity in vertebrates (McQueen and Parham 2002), i.e. a pattern modulated by the pathogen environment. If this is the case, selection acting on immune function would have been the driving force for maintaining an interesting form of alternative somatic diversification in the immune repertoire.

AUTHORSHIP

DB did all the expression experiments and the analysis of the duplicated exons. SM and TL built the phylogenies, KM, JA and KT and IC cloned the gene in *D. magna*. TF did the statistical analysis. LDP designed the experiments and wrote the paper together with DE, DB and SM.

AKNOWLEDGMENTS

We thank Brigitte Aeschbach for technical assistance and Dietmar Schmucker for support and helpful discussions. The *D.pulex* sequence data were produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) in collaboration with the Daphnia Genomics Consortium http://daphnia.cgb.indiana.edu.

D.B. is supported by the Portuguese Science Foundation (FCT). D. E. and I. C. were supported by the Swiss National Founds.

REFERENCES

Agarwala KL, Subramaniam G, Tsutsumi Y, Suzuki T, Kenji A, Yamakawa K. 2001. Cloning und Functional Characterization of DSCAML1, a Novel DSCAM-like Cell Adhesion Molecule that Mediates Homophilic Intercellular Adhesion. Biochem Bioph Res Co. 285:760-772.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. 25:3389-3402.

Bevilacqua A, Ceriani MC, Capaccioli SNA. 2003. Post-transcriptional regulation of gene expression by degradation of messenger RNAs. J Cell Physiol. 195:356-372.

Barrow A, Trowsdale J. 2006. You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signalling. Eur. J. Immunol. 36:1646 - 1653.

Celoto AM, Graveley B. 2001. Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. Genetics. 159:599-608.

Chen BE, Kondo M, Garnier A, Watson FL, Püettmann-Holgado R, Lamar DR, Schmucker D. 2006. The Molecular Diversity of Dscam Is Functionally Required for Neuronal Wiring Specificity in *Drosophila*. Cell. 125:607-620.

Colbourne JK, Hebert PDN. 1996. The systematics of north american *Daphnia* (Crustacean: Anomopoda): a molecular phylogenetic approach. Phil. Trans. R. Soc. Lond. B. 351:349-360.

Crayton M, Powell B, Vision T, Giddings M. 2006. Tracking the evolution of alternatively spliced exons within the Dscam family. BMC Evol Biol. 6:1-15.

Dong Y, Taylor HE, Dimopoulos G. 2006. AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biology. 4:e229.

Du Pasquier L. 2005. Diversify One Molecule to Serve Two Systems. Science. 309:1826-1827.

Ebert D, Rainey P, Embley TM, Scholz D. 1996. Development, life cycle, ultrastructure and phylogenetic position of *Pasteuria ramosa* Metchnikoff 1888: rediscovery of an obligate endoparasite of *Daphnia magna* Strauss. Phil. Trans. R. Soc. Lond. B. 351:1689-1701.

Fanning AS, Anderson JM. 1999. PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. J Clin Invest. 103:767-772.

Glenner H, Thomsen PF, Hebsgaard MB, Sørensen MV, Willerslev E. 2006. The origin of Insects. Science. 314:1183-1884.

Graveley B, Amardeep K, Dorian G, Lawrence ZS, Lee R, c. CJ. 2004. The organization and evolution of the Dipteran and Hymenopteran Down syndrome cell adhesion molecule (*Dscam*) genes. RNA.1499:1506.

Graveley BR. 2005. Mutually exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell. 123:65-73.

Hattori D, Demir E, Kim HW, Virahg E, S.L. Z, Dickson BJ. 2007. Dscam diversity is essential for neuronal wiring and self-recognition. Nature. 449:223-228.

Indik ZK, Park JG, Hunter S, Schreiber AD. 1995. Structure/function relationships of Fc gamma receptors in phagocytosis. Semin Immunol. 7:45-54.

Kleene R, Schachner M. 2004. Glycans and neural cell interactions. Nat Rev Neurosci. 5:195-208.

Kreahling JM, Graveley B. 2005. The iStem, a Long- Range RNA Seconday Structure Element Required for Efficient Exon Inclusion in the *Drosophila Dscam* Pre-mRNA. Mol Cell Biol. 25:10251-10260.

Lanier LL. 2001. Face off - the interplay between activating and inhibitory immune receptors. Curr. Opin. Immunol. . 13:326-331.

McQueen KL, Parham P. 2002. Variable receptors controlling activation and inhibition of NK cells. Curr. Opin. Immunol. 14:615-621.

Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J-h, Walz T, Wang J-h, Schmucker D. 2007. Structural basis of Dscam isoform specificity. Nature. 449.

Millard SS, Flanagan JJ, Pappu KS, Wu W, Zipursky L. 2007. Dscam2 mediates axonal tiling in the *Drosophila* visual system. Nature. 447:720-724.

Nei M, Hughes AL. 1992. Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: K. Tsuji, M. Aizawa, and T. Sasazuki, editors. Proceedings of the 11th Histocompatibility Workshop and Conference. Oxford: Oxford University Press. p. 27-38.

Nei M, Rogozin IB, Piontkivska H. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc. Natl. Sci. USA. 97:10866-10871.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121-152.

Neves G, Zucker J, Daly M, A C. 2004. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. Nat Genet.240-246.

Ota T, Nei M. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 11:613-619.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zypursky Sl. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell. 101:671-684.

Shapiro L, Love J, Colman DR. 2007. Adhesion molecules in the nervous system:structural insights into function and diversity. Annu Rev Neurosci. 30:451-474.

Sheng M, Sala C. 2001. PDZ domains and the organization of supramolecular complexes. Annu Rev Neurosci. 24:1-29.

Staub E, Rosenthal A, Hinzmann B. 2004. Systematic identification of immunoreceptor tyrosine-based inhibitory motifs in the human proteome. Cell Signal. 16:435-456.

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

Watson LF, Püttmann-Holgado FT, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, Schmucker D. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science. 309:1874-1878

Wojtowicz WM, Flanagan JJ, S.L. Z, Clemens J. 2004. Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific binding. Cell. 118:619-633.

Yamakawa K, Huo Y-K, Haendel MA, Hubert R, Chen X-N, Lyons GE, Korenberg JR. 1998. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. Hum Mol Genet. 7:227-237.

Zhang J, Rosenbergdagger HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Sci. USA. 95:3708-3713.

Millard, S. S., J. J. Flanagan, K. S. Pappu, W. Wu, and S. L. Zipursky. 2007. Dscam2 mediates axonal tiling in the Drosophila visual system. Nature **447**:720-U714.

Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. l. Zypursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity Cell **101**:671-684.

SUPPLEMENTARY MATERIAL

MATERIAL AND METHODS

Phosmid Libray The DNA to be use in the fosmid library was prepared in the following way: five hundred adult individuals (ca 1 gram of wet tissue) were kept in filtered culture medium with 50mg/L of Ampicillin (to reduce bacterial contamination) and 300 mg/L of Sephadex G-25 beads (Sigma-Aldritch) (to replace gut content). The culture medium was renewed every day for one week. This treatment was aimed at reducing the bacterial load and subsequent contamination of the fosmid library. The individuals were then harvested and frozen at - 20°C until DNA extraction. Genomic DNA was extracted from 2 grams of *Daphnia magna* (clonal line Mu11) using the Qiagen genomic tip protocol. Fosmid libraries were generated using the Copy Control^{TM} Fosmid cloning Kit (Epicenter, Madison, WI) following the manufacture's protocol. Briefly, 20 ug of genomic DNA was end-repaired and size fractionated in a pulse field gel with 1% SeaKem Gold Agarose (Cambrex Bio Science, Rockland ME) in 0.5X TBE buffer. DNA in the size range of 35 to 50 Kb was isolated by GELase treatment and the product was ligated into the vector pCC2FOS^{TM}. Ligations were transformed into T1-resistant *E. coli* cells (EPI300^{TM}-T1^{R}) by electroporation.

After quality control analysis of library, fosmid clones were picked to approximately 5X coverage on a Q-bot (Genetix, Newmilton, UK) and stored as individual clones grown in 384 well plates at -80 °C. To screen these clones for fosmids containing the gene of interest, pooled fosmids were screened with primers fn35f-r (seq) and IG1f-r (seq) designed to target exons near the 5' and 3' ends of the gene. Five positive clones were identified and one of the clones (1F5) was found to be positive for both primer pairs. End sequencing of all positive clones confirmed the placement of these clones relative to the *D. pulex* draft genome and that fosmid 1F5 spanned the entire Dscam gene in *D. pulex*. The insert from fosmid 1F5 was isolated as a *Sma*I digestion product by gel electrophoresis and GELase digestion. The insert was subsequently randomly sheared on a GeneMachines HydroShear (Genomic Solutions, Ann Arbor, MI,) to an average size of 3Kb. Sheared DNA was then end-repaired and size selected by agarose gel electrophoresis and the products were blunt end cloned into *Sac*I digested Puc-18 vector treated with Calf Intestinal Phosphatase (New England Biolabs, Ipswich, MA). After ligation and transformation into One Shot, Genehogs electrocompetent cells (Invitrogen, Carlsbad, CA). A plate of 384 clones was picked and sequencing template was prepared by rolling circle amplification (GE Healthcare, Piscataway, NJ) before sequencing on an ABI 3130 (Foster City, CA) capillary DNA sequencer.

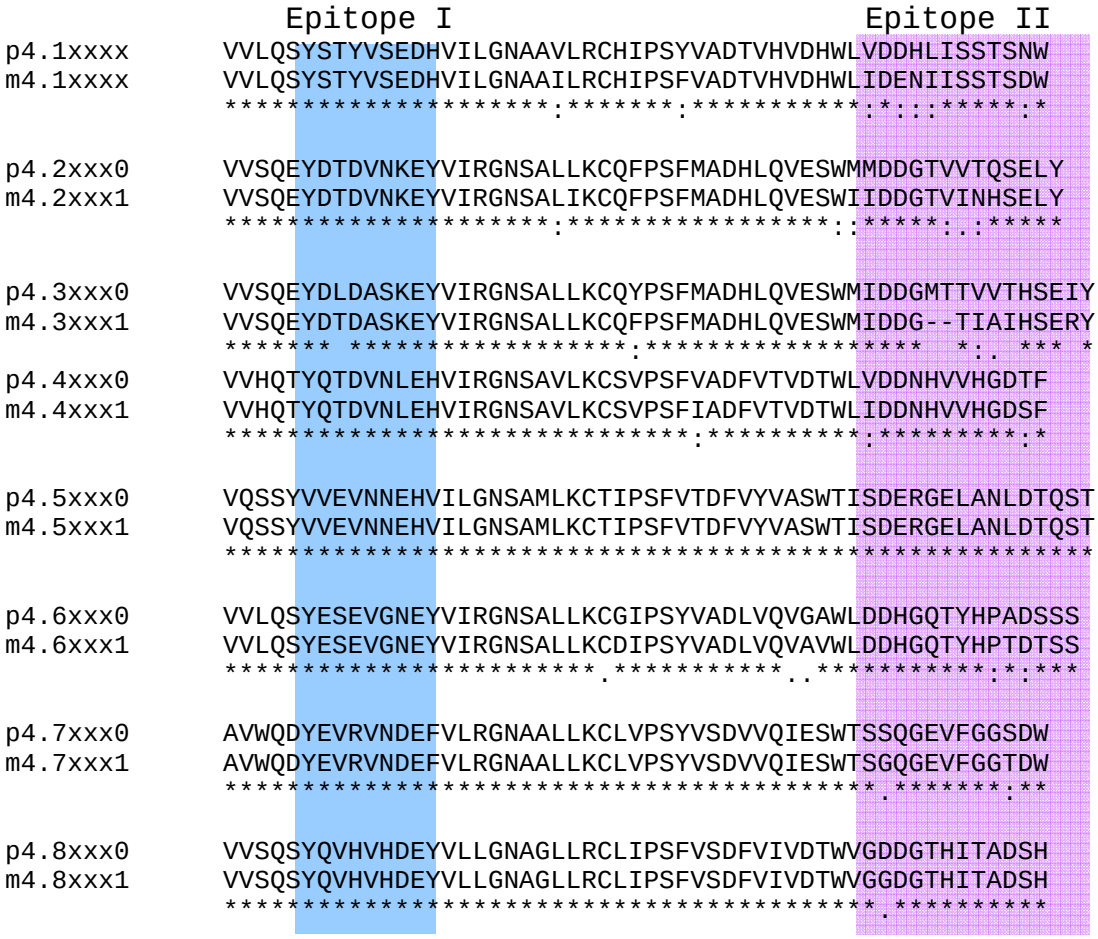
Accession numbers

human dscaml	aal57166.1
chimp dscaml	xp001158737.1
Dog dscam l	xp546506.2
Rat dscam l	xp236203.3
mouse dscaml	xp236203.3
zebrafish dscam	aat36313.1
chicken dscam	xp416734.2
opossum dscam	xp001370653.1
Dog dscam	xp544893.2
mouse dscam	np112451.1
Rat dscam	np598271
chimp dscam	Xp001171538.1
human dscam	aac17967.1
Flatworm (Fusaoka et al 2006)	Ab249988
Sea urchin	Xp793690
Bee dscaml	baf03050
Aedes dscaml	aael013409 pa
Dmel dscaml2	Cg32387
dmel dscaml1	c331190 pa
aedes dscam	aael010606
dmel dscam	aaf71926.1
tribolium dscam	Xp969935
Bee dscam	aat96374.1

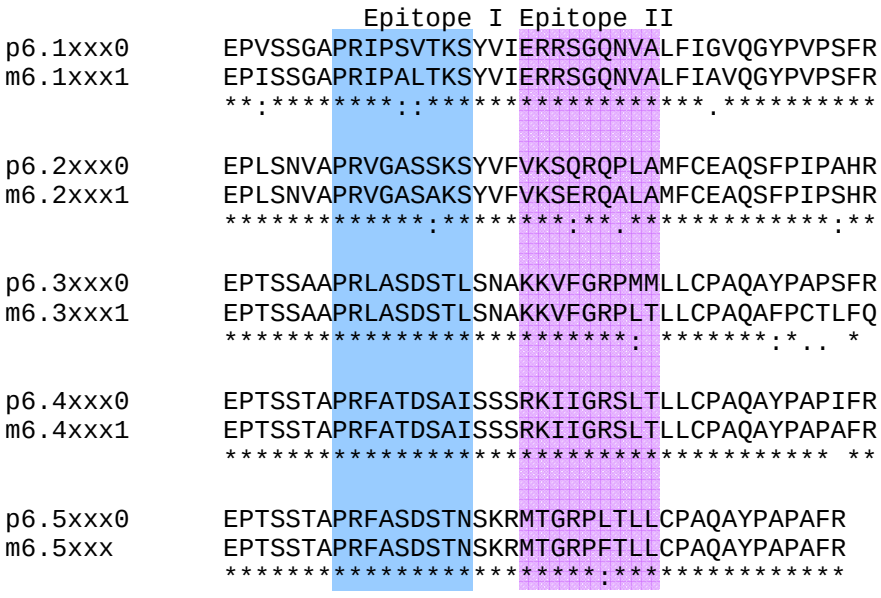
RESULTS AND DISCUSSION

Figure S1 Alignment of intervening sequences from array 6 in *D. magna*. In blue the reverse complementary sequence of the docking *Drosophila* consensus (Graveley 2005). In yellow putative segments corresponding to the selector sequences: Numbers on the left 1-2, 2-3 etc, refer to the intervening sequence position with respect to the exons, i.e. 1-2 refers to the intervening sequence between exons 6.1 and 6.2; Numbers 1), 23), 38) etc, refer to from which base of the intervening sequence the sequence is represented in the figure; Numbers on the right indicate the last base represented and/or the total number of bases in each intervening sequence. Intervening sequences have been grouped according to size.

Array 4



Array 6



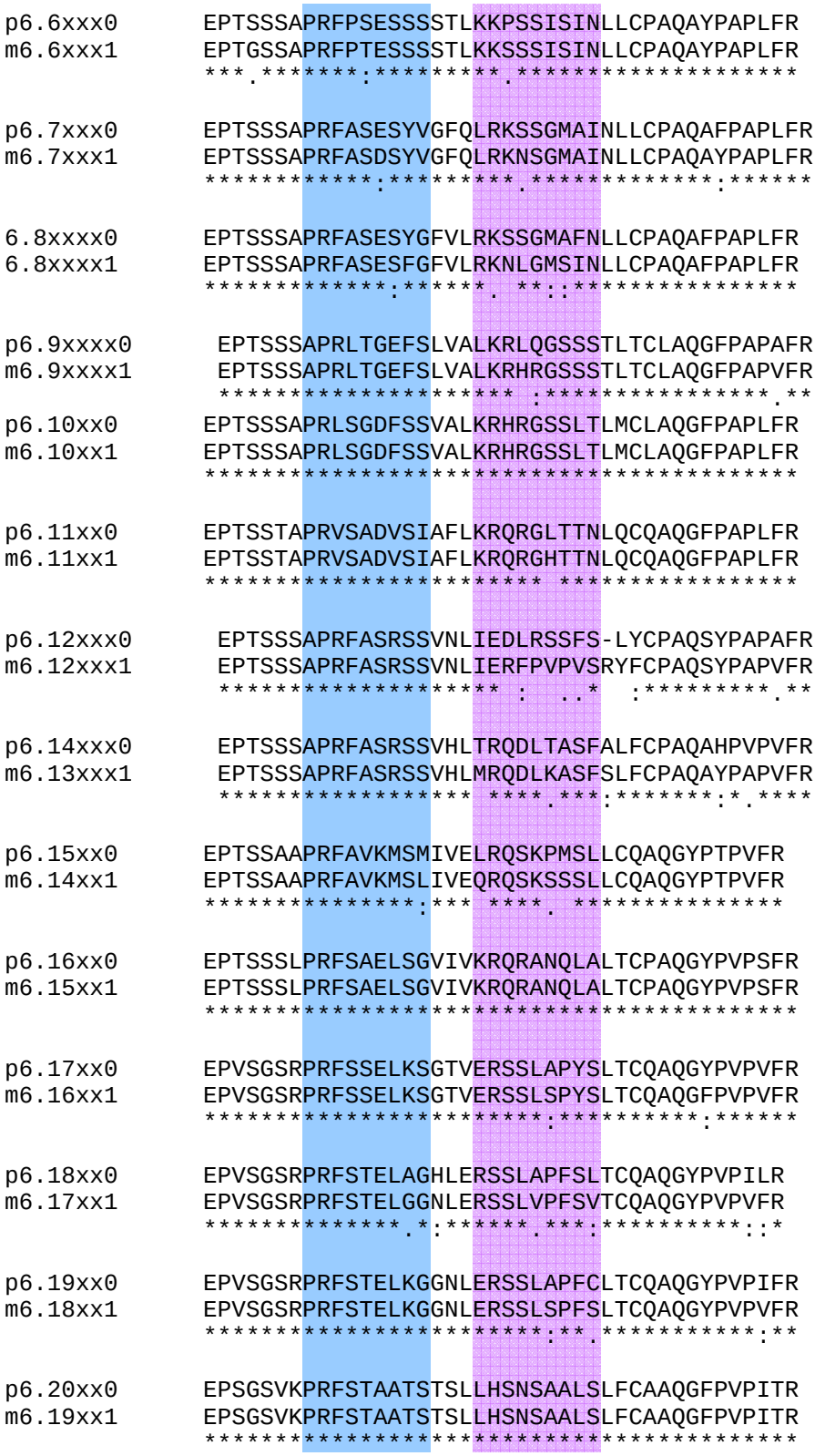

p6.6xxx0 EPTSSSAPRFPSESSSSTLKKPSSISINLLCPAQAYPAPLFR
m6.6xxx1 EPTGSSAPRFPTESSSSTLKKSSSISINLLCPAQAYPAPLFR
p6.7xxx0 EPTSSSAPRFASESYVGFQLRKSSGMAINLLCPAQAFPAPLFR
m6.7xxx1 EPTSSSAPRFASDSYVGFQLRKNSGMAINLLCPAQAYPAPLFR
6.8xxxx0 EPTSSSAPRFASESYGFVLRKSSGMAFNLLCPAQAFPAPLFR
6.8xxxx1 EPTSSSAPRFASESFGFVLRKNLGMSINLLCPAQAFPAPLFR
p6.9xxxx0 EPTSSSAPRLTGEFSLVALKRLQGSSSTLTCLAQGFPAPAFR
m6.9xxxx1 EPTSSSAPRLTGEFSLVALKRHRGSSSTLTCLAQGFPAPVFR
p6.10xx0 EPTSSSAPRLSGDFSSVALKRHRGSSLTLMCLAQGFPAPLFR
m6.10xx1 EPTSSSAPRLSGDFSSVALKRHRGSSLTLMCLAQGFPAPLFR
p6.11xx0 EPTSSTAPRVSADVSIAFLKRQRGLTTNLQCQAQGFPAPLFR
m6.11xx1 EPTSSTAPRVSADVSIAFLKRQRGHTTNLQCQAQGFPAPLFR
p6.12xxx0 EPTSSSAPRFASRSSVNLIEDLRSSFS-LYCPAQSYPAPAFR
m6.12xxx1 EPTSSSAPRFASRSSVNLIERFPVPVSRYFCPAQSYPAPVFR
p6.14xxx0 EPTSSSAPRFASRSSVHLTRQDLTASFALFCPAQAHPVPVFR
m6.13xxx1 EPTSSSAPRFASRSSVHLMRQDLKASFSLFCPAQAYPAPVFR
p6.15xx0 EPTSSAAPRFAVKMSMIVELRQSKPMSLLCQAQGYPTPVFR
m6.14xx1 EPTSSAAPRFAVKMSLIVEQRQSKSSSLLCQAQGYPTPVFR
p6.16xx0 EPTSSSLPRFSAELSGVIVKRQRANQLALTCPAQGYPVPSFR
m6.15xx1 EPTSSSLPRFSAELSGVIVKRQRANQLALTCPAQGYPVPSFR
p6.17xx0 EPVSGSRPRFSSELKSGTVERSSLAPYSLTCQAQGYPVPVFR
m6.16xx1 EPVSGSRPRFSSELKSGTVERSSLSPYSLTCQAQGFPVPVFR
p6.18xx0 EPVSGSRPRFSTELAGHLERSSLAPFSLTCQAQGYPVPILR
m6.17xx1 EPVSGSRPRFSTELGGNLERSSLVPFSVTCQAQGYPVPVFR
p6.19xx0 EPVSGSRPRFSTELKGGNLERSSLAPFCLTCQAQGYPVPIFR
m6.18xx1 EPVSGSRPRFSTELKGGNLERSSLSPFSLTCQAQGYPVPVFR
p6.20xx0 EPSGSVKPRFSTAATSTSLLHSNSAALSLFCAAQGFPVPITR
m6.19xx1 EPSGSVKPRFSTAATSTSLLHSNSAALSLFCAAQGFPVPITR

Figure S2 Amino acid alignment of orthologous exons from arrays 4, 6 of *D. pulex* (p) and *D. magna* (m). Symbols represent levels of amino acid identity between species: (*) full identity, (:) strongly similar, (.) weakly similar and ( ) no similarity. The boxes delimit Epitope I (blue box) and Epitope II (pink box) according to *D. melanogaster* (Meijers et al 2007).

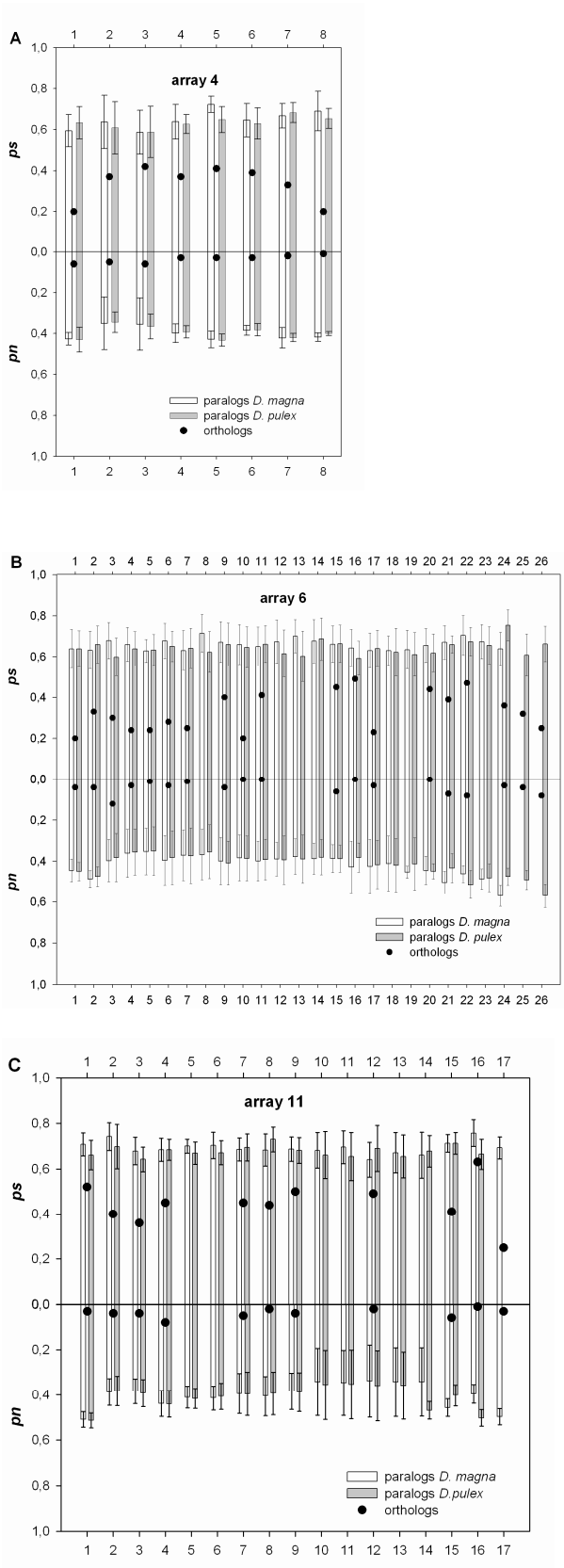


Figure S3 Number of synonymous (*ps*) and nonsynonymous substituions (*pn*) per synonymous and nonsynonymous sites respectively, of paralogs (bars) and orthologs (dots) for each Dscam array 4 (A), array 6 (B) and array (C). The bars represent the average *ps* and *pn* between paralogous exons within each cluster for both *Daphnia* species and the error bars its standard deviation. The dots represent the value of *ps* and *pn* for pairs of orthologous exons between the two *Daphnia* species identified by the Bayesian analysis and indicated on Fig.6b).

Array 4	**dn/ds**	**Array 6**	**dn/ds**	**Array 11**	**dn/ds**
4.1	0.26	6.1	0.18	11.1	0.03
4.2	0.10	6.2	0.09	11.2	0.07
4.3	0.10	6.3	0.34	11.3	0.08
4.4	0.06	6.4	0.11	11.4	0.12
4.5	0.05	6.5	0.03	11.5	na
4.6	0.00	6.6	0.11	11.6	
4.7	0.04	6.7	0.03	11.7	0.07
4.8	0.04	6.8	0.17	11.8	0.03
average	**0.08**	6.9	0.05	11.9	0.04
STDEV	**0.08**	6.10	0	11.10	na
		6.11	0	11.11	na
		6.12	na	11.12	0.02
		6.13	na	11.13	na
		6.14	0.13	11.14	na
		6.15	0.09	11.15	0.08
		6.16	0	11.16	0.00
		6.17	0.11	11.17	0.10
		6.18	0.22	Average	**0.06**
		6.19	na	STDEV	**0.04**
		6.20	0		
		6.21	0.13		
		6.22	0.11		
		6.23	na		
		6.24	0.06		
		6.25	0.1		
		6.26	0.28		
		average	**0.11**		
		STDEV	**0.09**		

Table S1 dn/ds of orthologous exons from arrays 4, 6 and 11 calculated by correcting *ps* and *pn* with the Jukes-Kantor formula (Ota and Nei 1994)**.**

	Dscam-hv	**Dscam**	**Dscam-L**
Daphnia magna, D.pulex	5	na	na
Drosophila melanogaster	6	na	11
Apis mellifera	4	na	12
Aedes aegypti	8	na	13
Danio rerio	na	17	na
Gallus gallus	na	17	na
Strongylocentrotus purpuratus	na	16	na
Dugesia japonica	na	19	na
Homo sapiens	na	na	15

Table S2. Number of glycosylation sites in variable and non variable Dscams determined with NetNglyc (http://www.cbs.dtu.dk/services/NetNGlyc/))

CHAPTER 2

EXPRESSION OF DSCAM IN THE CRUSTACEAN DAPHNIA MAGNA IN RESPONSE TO NATURAL PARASITES

Daniela Brites, Dieter Ebert and Louis du Pasquier

manuscript

ABSTRACT A vast diversity of isoforms of the Down syndrome cell adhesion molecule (Dscam) of insects and crustaceans is produced by mutually exclusive alternative splicing of dozens of internally tandem duplicated exons present in the Dscam locus. These exons code for segments or whole immunoglobulin domains of the protein. The diversity produced by alternative splicing plays a role in the development of the nervous system and it was suggested to be implicated in the immune defense of insects. In crustaceans like in insects, it has been shown to be expressed by immune cells. Here we tested whether the expression of Dscam is altered in the crustacean *Daphnia magna* challenged with several natural parasite species and strains. Furthermore we compared the repertoire of Dscam transcripts in nervous tissue and hemocytes in individuals infected or not with a naturally infective gram-positive bacterium. Hemocytes expressed lower transcript Dscam diversity in comparison with the nervous tissue. This shift was even more pronounced in hemocytes from infected *Daphnia*. However we found no effect of parasite infection on the usage of the alternative exons 4, or on the total amount of Dscam expressed. Yet, the finding of the same Dscam isoforms expressed in independent experiments suggests that associations between exons are functionally important.

INTRODUCTION

The highly diversified protein Dscam (Down syndrome cell adhesion molecule), already known for its essential role in the wiring of insect nervous system (Schmucker et al. 2000; Chen et al. 2006; Hattori et al. 2008), has been put forward as an exciting candidate for mediating specific immune responses in Arthropods (Kurtz and Armitage 2006). Much of

that is due to the fact that numerous different Dscam isoforms can be produced in hemocytes of one single individual by mutually exclusive alternative splicing of duplicated exons present in the Dscam locus (Neves et al. 2004; Watson et al. 2005). This has been reported initially in insects and later in crustaceans (Brites et al. 2008; Chou et al. 2009). Studies on *Drosophila melanogaster* (Watson et al. 2005) and *Anopheles gambiae* (Dong, Taylor, and Dimopoulos 2006) addressed in detail the function of Dscam in immunity and found support for it. However, not all evidences are in agreement (Vlisidou et al. 2009) and many important gaps need to be filled in order to have a sound understanding of the action of Dscam in immunity. Some of these gaps are difficult to address in model organisms such as *D. melanogaster*. Clonal reproduction and the use of natural endoparasites can help to shed light on some of these gaps. Here we study the expression of Dscam following infection of the asexual reproducing brachiopod crustacean *Daphnia magna* by several of its natural parasites. The gene Dscam encodes a protein composed extracellularly of immunoglobulin (Ig) and fibronectin III (FNIII) domains arranged in the following way, 9(Ig)-4(FNIII)-(Ig)-2(FNIII). Half of Ig2 and Ig3 domains and the entire Ig7, are coded by exons that are mutually exclusive alternatively spliced, while the other domains of the protein remain constant (Fig.1) The alternative exons are organized in 3 arrays in the Dscam locus (Fig.1). In insects and in the crustacean *Daphnia* the Dscam gene codes for isoforms that are membrane receptors with signaling capacity, although the intracellular domains in both groups differ in their motif organization (Schmucker et al. 2000; Brites et al. 2008).

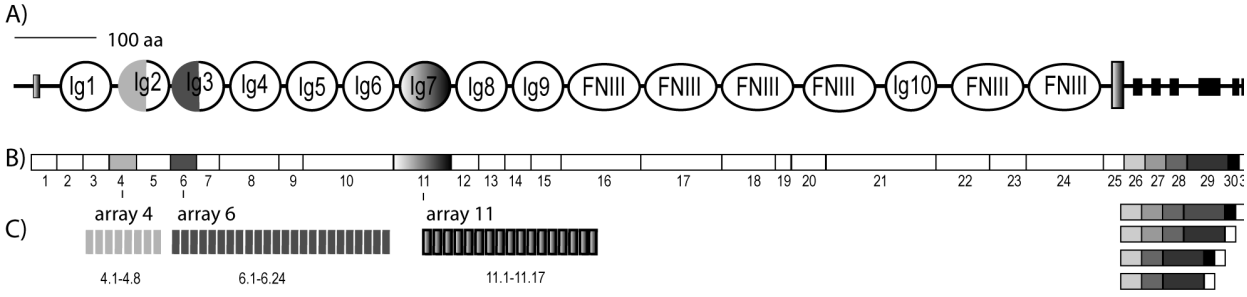


Figure 1 The Dscam of *D. magna* A) Protein domains; Ig-immunoglobulin domains; FNIII- fibronectin III domains. The grey and black boxes represent the transmembrane and cytoplasmic domains. B) mRNA, each box corresponds to a constitutive exons and the colored boxes 4,6 and 11, correspond to exons that are the result of mutual exclusive alternative splicing of arrays of duplicated exons which are present in three arrays, as indicated in C). Exons 26 to 31 code for alternative cytoplasmic tails (Brites *et al*, 2008). C) arrays of alternative exons 4, 6 and 11. Alternative cytplasmic tails following (Brites *et al*, 2008). Considering all splicing possibilities and alternative cytoplasmic tails *D. magna* can potentially produce 13056 different Dscam isoforms.

In *Daphnia*, alternative cytoplasmic tails are expressed, encoding either a tyrosine-based inhibition motif (ITIM) or an immunoreceptor tyrosine-based activation motif (ITAM), suggesting diversity in both recognition and effector capacities (Fig. 1) (Brites et al. 2008). Similarly, alternative cytoplasmic tails are expressed in *Drosophila* and an ITAM motif is also present in one of the alternative forms (Yu et al. 2009). In *Drosophila* and *Anopheles* Dscam is present in soluble forms produced by proteolytic cleavage in the hemolymph (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006). Interestingly, the Dscam of the decapod crustacean *Litopenaeus vannamei*, seems to code for isoforms that lack a cytoplasmic tail (Chou et al. 2009). Phagocytosis is an important cellular mechanism by which arthropods defend themselves from pathogens (Pham et al. 2007; Stuart and Ezekowitz 2008). It has been shown that knocking down Dscam by RNAi in third instar larvae of D. melanogaster and A. gambiae immune competent cells, reduces phagocytosis by approximately 45 to 60% (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006). Contrastingly, another study has shown that null Dscam mutant *D. melanogaster* embryonic hemocytes were still able to phagocyte bacteria as efficiently as their wild counterparts (Vlisidou et al. 2009). *Anopheles* mosquitos depleted of Dscam through gene silencing, suffered from high microbe proliferation in the hemolymph even in the absence of experimental challenge (Dong, Taylor, and Dimopoulos 2006). The same study has suggested that regulation of alternative splicing of exons belonging to array 4 seems to occur in Su5B cells, and to a lesser extent in adult mosquitos, in response to several pathogens. Finally, different Dscam isoforms have different binding affinities to bacteria (Watson et al. 2005) and in mosquito Su5B cells, isoforms induced by different pathogens had higher affinity for the inducer pathogen than for other pathogen species (Dong, Taylor, and Dimopoulos 2006).

We have previously shown that Dscam is expressed by hemocytes and nervous tissue in the crustacean *D. magna* (Brites et al. 2008). Its expression in hemocytes is not per se conclusive of its involvement in immunity given that at least in insects, but likely also in other invertebrates, hemocytes are multitasking cells involved, among other tasks, in developmental processes and wound healing (Vlisidou et al. 2009). Here we tested whether the expression of Dscam is modified quantitatively and qualitatively, following an infection by different natural parasites of *D. magna* by real time PCR quantification of both the total amount of Dscam transcript expression and the expression of the alternative exons from array 4. Natural *D. magna* populations exhibit highly specific responses (innate specific responses dependent on the genotype of the host and parasite) in relation to different parasite species and to different parasite strains (Carius, Little, and Ebert 2001; Vizoso, Lass, and Ebert 2005; Little, Kathryn, and Ebert 2006). We tested the effect

of infection by two microsporidia species (*Octosporea bayeri* and *Ordospora colligata*) and by two different isolates from the gram-positive bacterium *Pasteuria ramosa* on Dscam expression. Clonal lines of *D. magna* can be maintained in the laboratory by asexual reproduction allowing to study exactly the same host genotype under different parasite species/strains infections without confounding effects of germline polymorphisms. To evaluate the effect of infection in the usage of the three Dscam variable regions we characterized transcripts in hemocytes and compared it to the repertoire expressed in nervous tissue belonging to the same individuals exposed and unexposed to the bacteria *P. ramosa*.

MATERIAL AND METHODS

Host and parasite strains

The *D. magna* genotypes used were SP1-2-3, HO2 and Mu11 originally sampled in Finland, Hungary and Germany respectively. The parasites used were the microsporidia *Oc. bayeri* and *Or. colligata* and two different isolates of *P. ramosa* (P1 and P3). The host SP1-2-3 is susceptible to all parasites except for *P. ramosa* isolate P3 whereas HO2 and Mu11 are susceptible and resistant to *P. ramosa* P1, respectively. *Daphnia magna* genotypes were cloned in laboratory by propagating isofemale lines under constant light (light:dark cycle of 16:8 hours) and temperature conditions (20°). The lines were synchronized in a way that all individuals used in the experiments were born in the same day from mothers which had been raised under equal conditions for at least three asexual generations. None of the parasites used can be cultured *in vitro* and were thus grown in *D. magna* clones different from the ones used in the experiments.

Dscam expression assessed by real time quantitative PCR

RNA extraction and cDNA synthesis RNA was extracted using Trizol (INVITROGEN) following the manufacturer instructions and using 5 µg of RNAse free glycogen (INVITROGEN) to increase RNA yield. The final RNA pellet was dissolved in 20 µl RNAse free water and stored at -80 °C. Removal of genomic DNA and cDNA synthesis were done with the QuantiTect Reverse Transcription kit (QIAGEN) following the manufacturer instructions. The primers used in the kit above mentioned are a mix of oligo-dt and random primers.

Dscam relative quantification by quantitative real time PCR Expression was accessed by quantitative real time PCR using TaqMAN chemistry (AB Applied Biosystems) and the Applied Biosystems 7500 Fast Real-Time PCR system. Dscam expression was evaluated by quantifying all alternative exons 4 except for the exon 4.7 for which we did not

obtain specific amplification. The expression the housekeeping gene (ß-actin) was used to standardize all quantitative PCR measurements. The expression of the alternative exons was furthermore standardized by the expression of a constant Dscam region (exon 5) by dividing the relative expression values of each exon in each sample by the relative expression of exon 5 in the same sample. The amount of primers and probes used was optimized before the analysis and all fragments amplified had approximately 100 bp to ensure similar amplification efficiency between target and reference genes (primers and probes designed available in Tab. S1). All PCR reactions were replicated three times, and expression was quantified by using the $2^{-\Delta\Delta Ct}$ method (Kenneth and Thomas 2001). After PCR quantification all samples were run on a gel to ensure that specific amplifications were quantified. Three independent replicates per treatment combination were analyzed. We fitted the Dscam expression data to several general linear models (GML) for each of experiment done (Figures 2, 3 and 4). The response variable (relative expression) was log-transformed to ensure that residuals were normally distributed.

Experimental design Several experiments were done to compare the expression of Dscam in *D. magna* individuals exposed and unexposed to parasites. Each replicate in all experiments was composed of 10 individual Daphnia, five days old, placed together in 40 ml Daphnia artificial medium (ADAM) (Klüttgen et al. 1994; Ebert, Zschokke-Rohringer, and Carius 1998). Three replicates per treatment and control were used for PCR quantification and three other replicates per treatment were used to estimate the rates of infection. In the latter case, individuals were left until infections could be detected by eye, and in uncertain cases microscopically (Jensen et al. 2006). All parasite treatments were done by adding a suspension of spores of each parasite or of several parasites together depending on the experiment (see below). The control treatments were left unexposed, but otherwise treated in the same way. Animals were fixed in RNAlater (AMBION) and left overnight at 4°, after which they were dry-ice frozen in order to facilitate the dissection of the head. This was done in order to minimize the contribution of Dscam by the nervous system of the animal

Experiment 1- Expression of alternative exons 4 in resistant and susceptible D. magna hosts exposed to P. ramosa.

Six replicates (each with 10 individuals) of *D. magna* clone HO2 and six replicates of *D. magna* clone Mu11 were exposed to *P. ramosa* isolate P1. Controls for each genotype were replicated three times. Infections were done with a suspension of 10^6 parasite spores per replicate (10^5 spores per *D. magna* individual). At the time of this experiment it was unknown how long it takes for infections to take place and how long the host takes to mount an immune response. Infections can be detected

microscopically approximately one week after exposure (Ebert et al. 1996) and we chose this time point to evaluate Dscam expression under infection by P1. Seven days after exposure animals of three replicates per treatment were collected for RNA extraction. The three other replicates of each exposed *D. magna* genotype were changed to fresh medium and were used to assess the infection success of the parasites.

Experiment 2– Timing of Dscam expression during infection by three parasites.

Experiment 2 was set subsequently to assess Dscam expression over several days post-exposure to a mixture of the parasites *P. ramosa* (P1), *Oc. bayeri* and *Or. colligata*. The host genotype used in this case was SP1-2-3, which is susceptible to all parasites used. Here we hypothesized that if there is a change of the Dscam alternative exons repertoire in response to infection that should be associated with an up-regulation of the whole gene. Thus, only the constant exon 5 was used to quantify constitutive Dscam expression under infection. Exposures were done consecutively at 0, 20 and 40 hours by adding parasite spore mixtures to the medium containing $5x10^4$ spores per parasite per *D. magna* individual. *Daphnia magna* individuals from three replicates were collected at time 0 (before exposure), 2, 4, 6, 8, 10 and 13 days after the first exposure, both from the parasite exposed and unexposed treatments.

Experiment 3- Specificity of Dscam expression during infection by different parasites.

This experiment was identical to experiment 2 except that infection treatments were done by adding separately *P. ramosa* isolates P1 and P3 (to which SP1-2-3 is resistant) and *Oc. bayeri*. As described previously, parasite spores were released in a 0, 20 and 40 hours period but 10^5 spores per individual were used.

Expression of Dscam variability in the immune and nervous tissues assessed by cDNA sequencing

The associations between alternative exons from each array per Dscam molecule in brain and hemocytes of both infected and control individuals, were assessed by sequencing amplicons containing the three variable exons which had been obtained by RT-PCR. In two independent experiments (see below) hemocytes and brains from 15 individuals from one replicate of exposed and control groups were collected for subsequent RNA extraction. In both groups, hemolymph was withdrawn by capillary action upon introducing a twice pulled microcapillary glass tube (Harvard apparatus GC100TF-10) into the heart chamber. The hemolymph from 15 individuals was pooled and transferred to 50 µl of Daphnia cell culture medium without antibiotics (Robinson et al. 2006) and 2 µl were used for counting the number of cells using a THOMA counting

chamber to ensure that there were enough hemocytes for RNA extraction (only done in experiment 5, see below). Cells were then spun at 4000 rpm for 2 min, the buffer was removed and the pellet was immediately stored in dry ice. The remaining tissue of the individuals from which the hemocytes were withdrawn was stored in RNA later (AMBION) as described before. Their heads were cut and used for RNA extraction of brain sample. mRNA from hemocytes and brains was obtain with Dynalbeads technology (Dynalbeads mRNA Direct^{tm} Micro kit) following the manufacturer's instructions and the final RNA was eluted in 15 µl of RNAse free water. Reverse transcription and PCR, which were done in only one reaction with OneStep RT-PCR Kit (QUIAGEN) following the manufacturer's instructions, using approximately 0.02 µg of RNA in both hemocytes and brain obtained from infected and uninfected individulas and Dscam specific primers (forward primer ATCGTCTCCGCAGACATCC; reverse primer TGCCTTGTCTGTAGGTTCGAC). The following RT-PCR program was used: 30 min. at 50°, 15 min at 95° followed by 40 cycles with denaturing at 94° during 30 sec, annealing at 57° during 30 sec and extension at 72° during 2 min and a final extension step of 10 min at 72°. The resultant amplicon had 1.9 kb and included variable exons from arrays 3, 6 and 11. The PCR products were cloned in a pCR 2.1- TOPO vector (INVITROGEN) and sequenced using the M13 reverse and forward primers.

Experiment 4 – Expression of all three Dscam arrays, in later stages of infection by P.ramosa.

At the same time that experiment 1 described above was set, additional replicates of infected (2 replicates) and uninfected (2 replicates) composed each of 15 *D. magna* (H02) individuals, were assigned for assessing the expression of the three variable arrays. The animals were collected at a later stage of infection by *P. ramosa* isolate P1 (30 days) and hemocytes and brains obtained from the same individuals were used for RNA extraction. We succeeded in obtaining Dscam amplification for hemocytes in only one of the infected replicates and in none of the control replicates. For that reason no expression of control animals could be analyzed. The PCR fragments containing transcripts from nervous tissue and hemocytes were cloned as described and twenty-five transformants per tissue sampled were sequenced.

Experiment 5 - Expression of all three Dscam arrays at 2 day post-exposure to P. ramosa.

In this experiment nine groups of 15 females of 22 days old *D. magna* (SP1-2-3) individuals, were kept in 40 ml ADAM. Three groups were left unexposed and the rest were exposed twice to *P. ramosa* isolate P1 within 40 hours. The parasite doses used were 10^4 spores per individual *Daphnia* in the first exposure and 10^5 in the second. Forty eight hours after the first

Experiment	*D. magna* Genotype	Parasite species/strains	Sampling (days)*	RNA origin	Dscam region targeted	Figures &Tables
1	HO2 (susceptible) Mu11 (resistant)	*P. ramosa* P1	7	Whole body without head	Exons 4, except 4.7	Fig. 2
2	SP1-2-3 (susceptible)	Mixture of *Oc. bayeri*, *Or. coligata* *P. ramosa* P1	2, 4, 6, 8, 10, 13	Whole body without head	Exon5	Fig. 3
3	SP1-2-3 (resistant to P3, otherwise susceptible)	*Oc. bayeri*, *P. ramosa* P1 and P3	2, 4, 6, 8, 10, 13	Whole body without head	Exon5	Fig. 4
4	HO2 (susceptible)	*P. ramosa* P1	30	Hemocytes and brain	Transcripts with Ig2 to Ig7 coding exons	Fig. 6
5	SP1-2-3 (susceptible)	*P. ramosa* P1	2	Hemocytes and brain	Transcripts with Ig2 to Ig7 coding exons	Fig. 5 Table2

Table 1 Overview of the five experiments. *days after the first exposure

exposure, hemocytes and brains from 15 individuals from the unexposed and from three of the exposed groups were collected for subsequent RNA extraction. The animals of the other remaining three replicates were changed to new medium and used to assess infections rate. Hemocytes were count to ensure amplification from both infected and uninfected individuals. Nevertheless, we obtained Dscam RNA from hemocytes in only one exposed and unexposed replicates. We used cDNA of brain samples and hemocytes belonging to the same individuals to obtain and clone PCR fragments as described above. Fifty transformants per tissue and treatment were sequenced.

Estimating Dscam transcript diversity The sequence data obtained from the experiments described was used to estimate several diversity indices using EstimatesS version 8.2 (Colwell 2006). Transcript diversity was calculated using the Simpson and Shannon indices.

The Shannon index (D) was furthermore used to estimate evenness (E) in the following way $E=e^{D}/N$ where N is the total number of different isoform sequences in the sample. The percentage of coverage achieved by our sampling was calculated by Good's method using the number of singletons *n* (transcripts that occurred only once in a certain sample) in the following way, $(1-n/N) \times 100$ (Good 1953).

RESULTS

Experiments 1, 2 and 3

An overview of all experiments and their specificities is given in Table 1. We found no significant differences in Dscam expression level between exposed individuals and controls in experiment 1, 2 and 3 (Fig. 2, 3 & 4). In experiment 1, the only significant effect found in Dscam expression was between exons (Fig. 2). Exons 4.4, 4.5 and 4.6, independently of the *D. magna* genotype or parasite infection, were significantly less expressed than the remaining exons (Fig. 2, for the three cases $p \leq 0.006$). In experiment 2, the expression of Dscam on day 2 of sampling was significantly higher than in the other days (Fig. 3, p=0.02). However, testing three parasites one by one, did not reveal a treatment effect (experiment 3, Fig. 4).

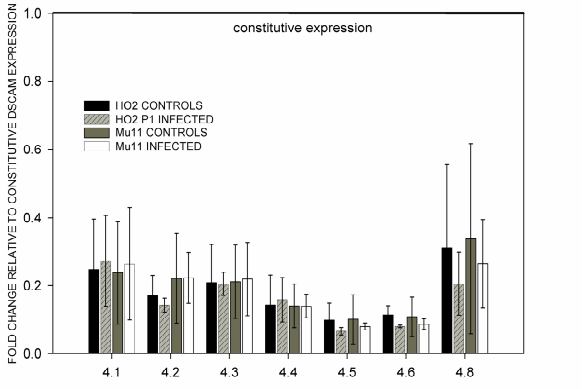


Figure 2 Relative expression of Dscam alternative exons from array 4 presented as fold change relative to the constitutive levels of Dscam produced (1) in susceptible (HO2) and resistant hosts (Mu11), 7 days exposed or not (controls) to the gram-positive bacteria *P. ramosa* (experiment 1). Each bar corresponds to the mean of three independent replicates and the error bars represent standard deviations. Dscam relative expression (RE) was fitted to the GML model log(RE)= genotype+exposure+exon+genotype:exposure. We found no statistical significant effect of parasite exposure (F=0.26, p=0.59), or of *D. magna* genotype (F=0.28, p=0.6) or of an interaction between both. Expression is significantly different between exons (F= 11.39, p<0.001).

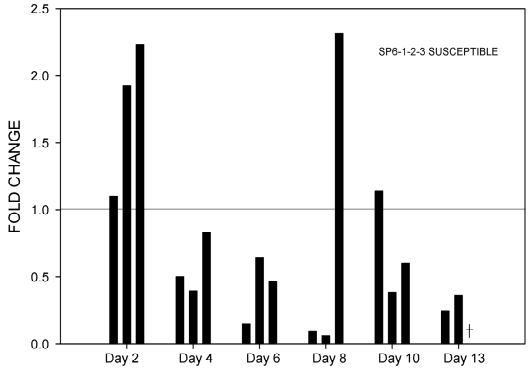


Figure 3 Relative expression of total Dscam (exon 5) of exposed SP1-2-3 individuals in relation to controls (Baseline) during several days post-exposure to a mixture of the microsporidia parasites (*O. bayeri* and *Or. colligata*) and the gram-positive bacteria *P. ramosa* (experiment 2). Three independent replicates per day post-exposure are depicted. Dscam relative expression (RE) was fitted to the GML model log(RE)=days+exposure+days:exposure. The only significant effect found was for day 2 (F=2.87, p=0.008) (exposure, F=0.75, p=0.39; interaction between exposure and day of sampling, F=0.5, p=0.76).

The infections in the susceptible hosts were always 100% successful in the replicates of the experiment that were used to assess infection rates. Thus, the animals used for testing Dscam expression were most likely infect as well. As expected, none of the exposed resistant host genotypes developed an infection.

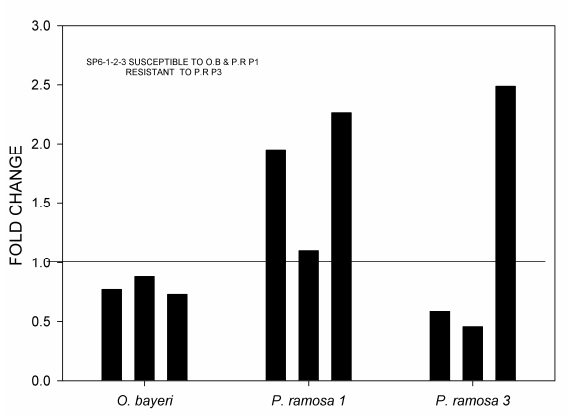


Figure 4 Relative expression of total Dscam of exposed SP1-2-3 individuals in relation to controls (baseline), 2 days post-exposure to the microsporidia parasite *O. bayeri* and to two isolates of and the gram-positive bacteria *P. ramosa* (experiment 3) . The infections by *O. bayeri* and *P. ramosa* P1 were 100% successful and no individual was infected by *P. ramosa* P3. Three independent replicates per are depicted. Dscam relative expression (RE) was fitted to the GML model log(RE)=exposure+parasite. No significant effects were found (exposure, F=0.02, p=0.8; parasite, F=0.9, p=0.48)

Experiments 4 & 5

Transcripts containing the three variable regions were obtained from nervous tissue and hemocytes from the same infected individuals, 30 days after exposure to *P. ramosa* (experiment 4) and from controls and exposed individuals, 2 days after exposure (experiment 5). We will mostly discuss the results obtained from exposed and control treatments from experiment 5. Experiment 4, from which we have no controls, will be mainly discussed in comparison with a similar experiment done previously (Brites et al. 2008). In both experiments, we used identical amounts of RNA from all treatments for performing the one-step RT-PCR, nevertheless the nervous tissue yielded more cDNA (Fig. 5A, 6A). The expressed diversity of arrays 4 and 6, but not of array 11, tends be higher in the brain than in hemocytes (Table 2). Comparing the diversity of hemocytes between infected and uninfected individuals revealed only a small effect on array 6 (Table 2).

Dscam	**control**		**infected**	
region	brain	hemocytes	brain	hemocytes
Array 4	18	14	19	17
Array 6	38	31	44	23
Array 11	25	25	29	28

Table 2 Expressed array diversity of exons calculated as the number of different exons found in each array per treatment divided by the total number of exons expressed in each array in control and infected individuals (%) (experiment 5).

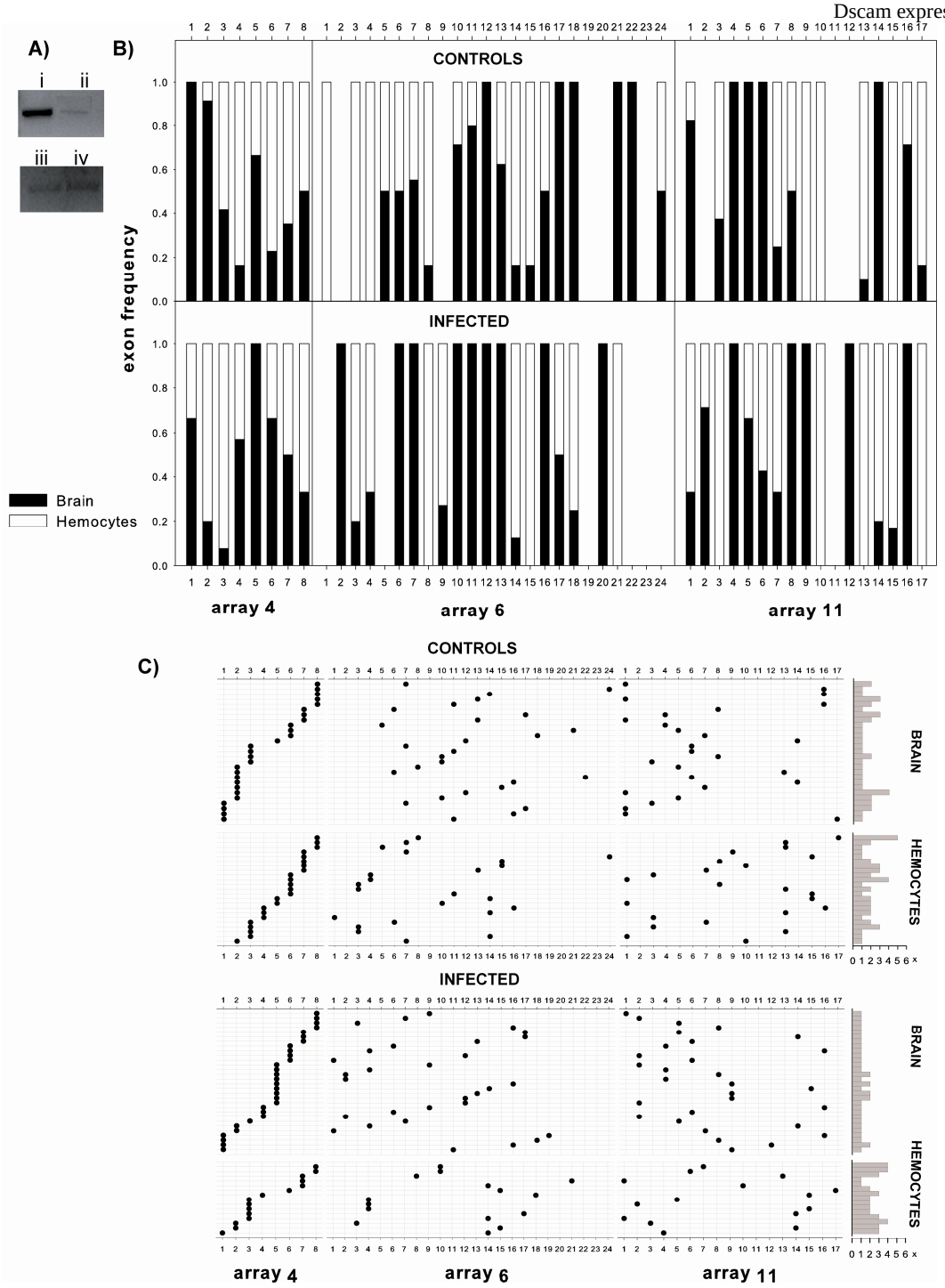


Figure 5 Experiment 5 A) *Daphnia magna* expression of a Dscam region containing the variable exons coding for Ig2, Ig3 and Ig7 (1850 bp) in brain and hemocytes of the same exposed and unexposed individuals, 2 days after exposure to *P. ramosa* P1. I - Controls brain; II -exposed brain; III - control hemocytes and IV - exposed hemocytes. The number of estimated hemocytes from which RNA was extracted was approximately 37 x 10^3 and 10^4 from control and exposed individuals respectively B) Exon usage frequency in brains and hemocytes from the same individuals. Bars correspond to the usage of each exon in brain and hemocytes relative to the total number of the times the exon was observed in the same individuals. C) Association of exons from each array in single mRNA molecule from brain and hemocytes belonging to the same individuals. The bars on the right side of the graph represent the absolute number of times each association was observed. Number of transcripts sequenced: brain control n=42; hemocytes control= 45; brain infected=35; hemocytes infected=39.

When examining how exons from each array associate with each other in forming the mRNA, a remarkable difference between hemocytes and brain emerged. Using various indicators of diversity, the brain expressed a higher total diversity of Dscam transcripts than hemocytes (Fig. 5C, Tab. 3). Hemocytes expressed a lower total diversity of transcripts and on average more of each one as shown by the lower evenness estimates (an evenness of 1 in a given sample would mean that all different transcripts would be present only once in that sample). Differences in abundance of transcripts have to be taken carefully though

Estimates	**Experiment 4**		**Experiment 5**			
	Hemocytes	**Brain**	**Hemocytes**		**Brains**	
	Infected N=17	**Infected N=21**	**Controls N=45**	**Infected N=39**	**Control N=42**	**Infected N=35**
singletons	5	17	9	2	17	25
Shannon's diversity index	2.91	9.03	2.96	2.64	3.32	3.21
Simpson's diversity index	15.11	105	26.72	19.5	51.7	93
Evenness (D)	0.53	0.87	0.42	0.36	0.6	0.79
Good's estimator coverage %	71	19	80	94	59	28

Table 3 Estimations of transcript diversity and sequencing coverage

because they could be influenced by the number of PCR cycles. Given the low amplification yield obtained for hemocytes, we think that this effect was likely not very significant, but we cannot exclude it completely (Fig. 5A, Fig. 6A). Hemocytes of infected animals exhibited a further reduction in diversity in relation to hemocytes of uninfected animals (Fig. 5, Tab. 3). The Good's estimator of coverage is 80% and 94% for hemocytes from control and infected individuals, respectively. That indicates that only 20 and 6 additional transcripts would be expected respectively, if 100 additional transcripts would be sampled. The transcript sampling was much more incomplete in the case of the brain (Tab. 3).

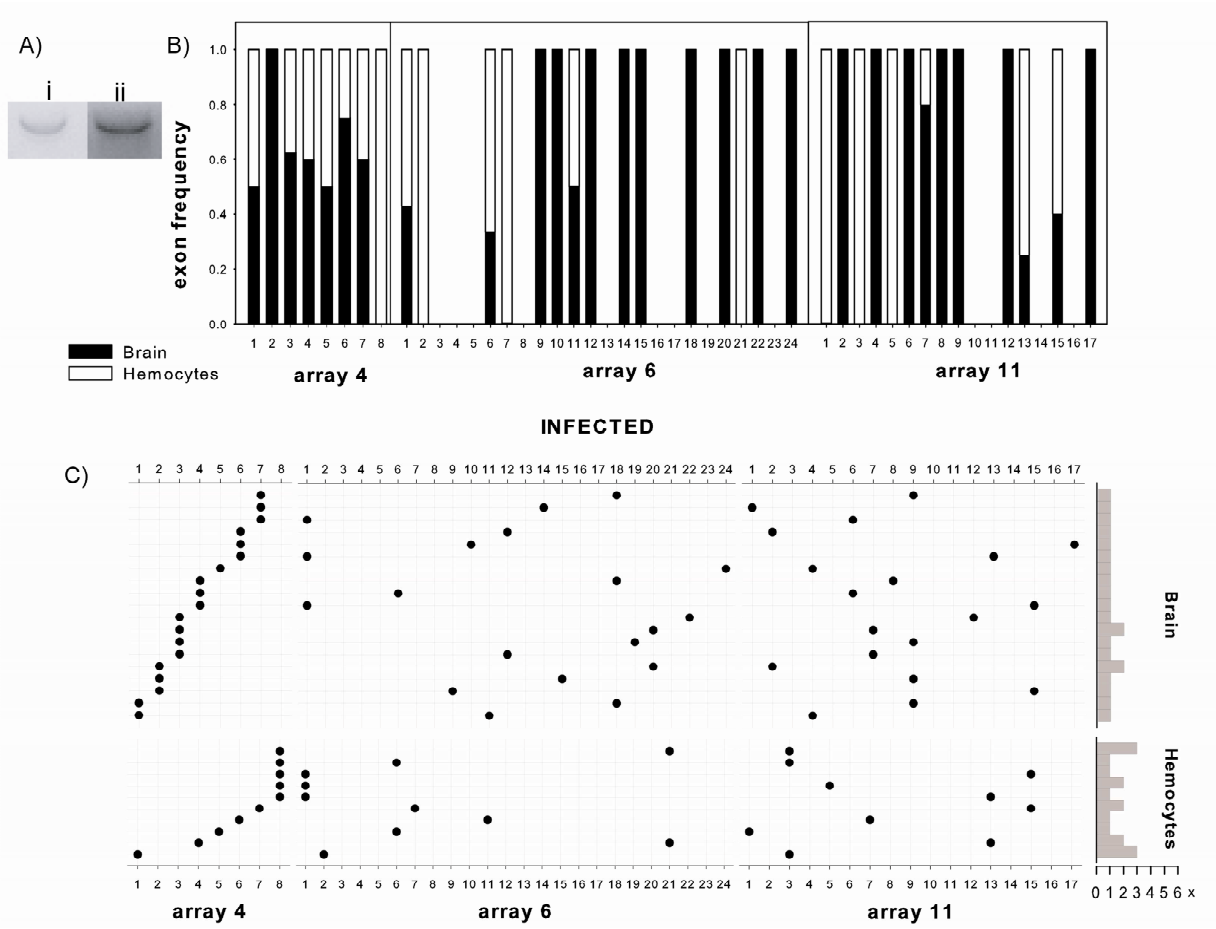


Figure 6 Experiment 4 A) *Daphnia magna* expression of a Dscam region containing the variable exons coding for Ig2, Ig3 and Ig7 (1850 bp) in brain and hemocytes of infected individuals with 30 days old infections by *P. ramosa* P1. RT-PCR was performed on RNA obtained from the brains and hemocytes of 15 cloned and synchronized *D. magna* HO2 individuals per treatment. I – infected hemocytes; II -infected brain. B) Exon usage frequency in brains and hemocytes from the same individuals. Bars correspond to the usage of each exon in brain and hemocytes relative to the total number of the times the exon was observed in the same individuals. C) Association of exons from each array in single mRNA molecule from brain and hemocytes belonging to the same individuals. The bars on the right side of the graph represent the absolute number of times each association was observed (brain infected, N=21; hemocytes infected, N=17).

In experiment 5, hemocytes from infected and uninfected individuals expressed different isoforms with the exception of isoform 4.3+6.14+11.1, which occurred once and three times in control and in infected hemocytes, respectively. Other transcripts, had common associations between exons from array 6 and 11 (Fig. 5C); the association between exon 6.3 and 11.13 occurs three and four times in control and infected hemocytes respectively, whereas it was never observed in the brain. The association between 4.7 and 6.13 was found twice in the nervous tissue from infected and uninfected individuals and never in hemocytes. The probability of finding any exon combinations several times in independent treatments can be roughly estimated by multiplying the probabilities of usage of one exon in each array (one mutually exclusive mutually spliced exon divided by the number of possible exons in that array). Under a random model (i.e. each exon on one array has the same chance to be incorporated in a transcript), the likelihood of finding twice, for instance, any combination of exons 6 and 11, would be 6 in 10^6 transcripts ($((1/24 \times 1/17)^2$). From each treatment 35 to 42 transcript sequences were obtained reducing that likelihood even further.

In experiment 4, the nervous tissue also exhibited higher transcript diversity and evenness than hemocytes (Fig. 6, Tab. 3). Common transcripts expressed by hemocytes were found between this and another experiment done previously under similar conditions, using the same *D. magna* genotype and *P. ramosa* isolate (Brites et al. 2008). We found transcript 4.8+6.1+11.15 once and five times respectively. In both experiments, exons 4.8 and 6.1 were often found associated, four and five times in the present and in the previous study (Brites et al. 2008), respectively. In this case, given that no control individuals were analyzed, it is not possible to discern whether that could be a consequence of infection.

DISCUSSION

The regulation of alternative exons from array 4 has been suggested to occur in both cell lines and adult mosquitos challenged with several pathogen species (Dong, Taylor, and Dimopoulos 2006). We tested whether that could be the case in the crustacean *D. magna*, using two genotypes that were either resistant or susceptible to a natural isolate of the gram-positive bacterium *P. ramosa* but did not find supporting evidence. That could be due to the fact that we missed the time when such effects might have taken place, or that Dscam is not involved in the resistance of *D. magna* to *P. ramosa*.

We hypothesized that if there is a change of the Dscam alternative exons repertoire in response to infection that should be associated with an increase in the expression of the whole gene and searched for up-regulation of Dscam under infection by other natural parasite species and throughout different post-exposure days.

However, we did not find up-regulation of Dscam neither in resistant nor in susceptible hosts. Despite the fact that cloned host lines of synchronized individuals were used in the experiments, the variation between replicates was high (Fig. 2-4). We can exclude PCR as a source of variation given that each PCR reaction was replicated three times and outlier measurements were removed, but whether the variation is biological or if it resides at the level of the RNA extraction and/or cDNA synthesis is unclear. To the absence of an effect could also contribute that in these experiments the whole body (without head) was used for RNA extraction. With this procedure, we could reduce the contribution of Dscam from the brain, but to which extent is unclear. Another possibility is that β-actin is not an adequate expression control gene, given that Dscam has been shown to interact with signaling proteins which are regulators of the actin-based cytoskeleton (Schmucker et al. 2000). Nevertheless, the work done by (Dong, Taylor, and Dimopoulos 2006) also reports an absence of up-regulation of the constitutive Dscam levels under infection, despite the significant effects of parasite challenge in modifying the expression of the alternative exons 4. This may be explained if the number of Dscam molecules present in cells is constant and only qualitative, but not quantitative changes in transcripts occur. Much remains to be done to find the mechanism of regulation of splicing in the context of an immune function.

Differences between nervous and immune Dscam repertoires may lie mainly in the associations between alternative exons and in the expressed amount of each isoform. We found that hemocytes expressed reduced repertoires but likely higher amounts of certain isoforms. Our results were obtained under homogeneous conditions, and in agreement with a previous study (Brites et al. 2008), in which however, hemocytes and brains belonged to animals of different genotype and different ages. This finding is consistent with an immune function of Dscam in hemocytes. Each individual isoform being present in higher concentrations would increase its functional specific capacities to bind to antigens (Brites et al. 2008).

Some expressed associations of exons were found to be common between independent treatments and experiments, mainly in hemocytes and in a lower extent in the brain. The likelihood of finding the same associations in different experiments by chance is low. Thus, the uneven expression of certain exon combinations may be determined by challenges rather than governed by chance. Several lines of evidences on how splicing is regulated in arrays 4 and 6, suggest that the regulatory sequences involved in splicing of each array are not the same, implying that the regulation of splicing of each array is independent of the other arrays (Graveley 2005; Kreahling and Graveley 2005; Olson et al. 2007). However, if certain associations between exons are important, it is possible that a further level of regulation acting

simultaneously in more than one array comes into place. Our results encourage new experiments evaluating transcription of the three variable Dscam regions in different tissues and under different parasites challenges.

Our results suggest furthermore, that if there is a role of Dscam in *D. magna* in response to the natural parasites tested, the effect is probably not very strong. We experienced repeatedly difficulties in obtained Dscam mRNA from hemocytes in comparison to whole bodies or brain suggesting that hemocytes express low amounts of Dscam in *D. magna*.

We consider that at this point it is still not possible to rule out the possibility that the role of Dscam in immunity is secondary, and that the main function of the different isoforms in hemocytes is, perhaps in a somehow similar way to what happens in the interactions between neurons, to provide them with a self-recognition system. This would prevent the formation of cell aggregation, allowing circulation in the hemolymph following the same mechanisms proposed for nervous cells (for a review see, Hughes et al. 2007 and Hattori et al. 2008). Under this scenario, immune related phenomena, such as lower phagocytosis rate and reduced survival as a consequence of Dscam knock-down (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006) could perhaps be a side-effect of a deficient population of hemocytes acting synergistically with parasite challenges. The existence of soluble circulating isoforms and the reduced transcript repertoires expressed by hemocytes are however, not fully consistent with this hypothesis. Moreover, structural and molecular evolution aspects of the variable regions of Ig2 and Ig3 suggest that Dscam could be involved in direct recognition of antigens (Meijers et al. 2007; Brites et al. 2010). A clear understanding of these aspects is necessary for a comprehensive view of how Dscam could contribute to explain immune phenomena such as immune priming or specificity of certain immune functions in insects and crustaceans (Kurtz and Franz 2003; Sadd and Schmid-Hempel 2006; Roth and Kurtz 2009).

AKNOWLEDGMENTS

We thank Dietmar Schmucker for support and helpful discussions.

REFERENCES

Brites, D., F. Encinas-Viso, D. Ebert, L. Du Pasquier, and C. R. Haag. 2010. Signatures of selection on duplicated alternatively spliced exons of the Dscam gene in *Daphnia* and *Drosophila*. *in preparation*.

Brites, D., S. McTaggart, K. Morris, J. Anderson, K. Thomas, I. Colson, T. Fabbro, T. J. Little, D. Ebert, and L. Du Pasquier. 2008. The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Molecular Biology and Evolution **25**:1429-1439.

Carius, H. J., T. Little, and D. Ebert. 2001. Genetic variation in a host-parasite association:potential for coevolution and frequency dependent selection. Evolution **55**:1136-1145.

Chen, B. E., M. Kondo, A. Garnier, F. L. Watson, R. Püettmann-Holgado, D. R. Lamar, and D. Schmucker. 2006. The Molecular Diversity of Dscam Is Functionally Required for Neuronal Wiring Specificity in Drosophila. Cell **125**:607-620.

Chou, P. H., H. S. Chang, I. T. Chen, H. Y. Lin, Y. M. Chen, H. L. Yang, and K. C. H. C. Wang.

2009. The putative invertebrate adaptive immune protein Litopenaeus vannamei Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. Developmental and Comparative Immunology **33**:1258-1267.

Colwell, R. K. 2006. EstimateS:statisticak estimation of species richness and shared species from samples. Version 8.

Dong, Y., H. E. Taylor, and G. Dimopoulos. 2006. AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biol **4**:e229-.

Ebert, D., P. Rainey, T. M. Embley, and D. Scholz. 1996. Development, life cycle, ultrastructure and phylogenetic position of Pasteuria ramosa Metchnikoff 1888: Rediscovery of an obligate endoparasite of Daphnia magna Straus. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **351**:1689-1701.

Ebert, D., C. D. Zschokke-Rohringer, and H. J. Carius. 1998. Within- and between-population variation for resistance of *Daphnia magna* to the bacterial endoparasite *Pasteuria ramosa*. Proc. R. Soc. Lond. B **265**:2127-2134.

Good, J. I. 1953. The population frequencies of species and the estimation of population parameters. Biometrika **40**:237-264.

Graveley, B. R. 2005. Mutually exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell **123**:65-73.

Hattori, D., S. S. Millard, W. M. Wojtowicz, and S. L. Zipursky. 2008. Dscam-Mediated Cell Recognition Regulates Neural Circuit Formation. Annual Review of Cell and Developmental Biology **24**:597-620.

Hughes, M. E., R. Bortnick, A. Tsubouchi, P. Baumer, M. Kondo, T. Uemura, and D. Schmucker. 2007. Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron **54**:417-427.

Jensen, K. H., T. Little, A. Skorping, and D. Ebert. 2006. Empirical support for optimal virulence in a castrating parasite. Plos Biology **4**:1265-1269.

Kenneth, L. J., and S. D. Thomas. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods **25**:402-408.

Klüttgen, B., U. Dülmer, M. Engels, and H. Ratte. 1994. ADaM, an artificial freshwater for the culture of zooplankton. Water Research:743-746.

Kreahling, J. M., and B. Graveley. 2005. The iStem, a Long- Range RNA Seconday Structure Element Required for Efficient Exon Inclusion in the *Drosophila Dscam* Pre-mRNA. Molecular and Celular Biology **25**:10251-10260.

Kurtz, J., and S. A. Armitage. 2006. Alternative adaptive immunity in invertebrates. Trends Immunol **27**:493-496.

Kurtz, J., and K. Franz. 2003. Evidence for memory in invertebrate immunity Nature **425**:37-38.

Little, T., W. Kathryn, and D. Ebert. 2006. Parasite-Host specificity:experimental studies on the base of parasite adaptation. Evolution **60**:31-38.

Meijers, R., R. Puettmann-Holgado, G. Skiniotis, J.-h. Liu, T. Walz, J.-h. Wang, and D. Schmucker. 2007. Structural basis of Dscam isoform specificity. Nature **449**:487-491.

Neves, G., J. Zucker, M. Daly, and C. A. 2004. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. Nature Genetics:240-246.

Olson, S., M. Blanchette, J. Park, Y. Savva, G. W. Yeo, J. M. Yeakley, D. C. Rio, and B. R. Graveley. 2007. A regulator of Dscam mutually exclusive splicing fidelity. Nature Structural & Molecular Biology **14**:1134-1140.

Pham, L. N., M. S. Dionne, M. Shirasu-Hiza, and D. S. Schneider. 2007. A specific primed immune response in Drosophila is dependent on phagocytes. PLoS Pathog **3**:e26.

Robinson, C. D., S. Lourido, S. P. Whelan, J. L. Dudycha, M. Lynch, and S. Isern. 2006. Viral transgenesis of embryonic cell cultures from the freshwater microcrustacean Daphnia. J Exp Zool A Comp Exp Biol **305**:62-67.

Roth, O., and J. Kurtz. 2009. Phagocytosis mediates specificity in the immune defence of an invertebrate, the woodlouse Porcellio scaber (Crustacea: Isopoda). Dev Comp Immunol **33**:1151-1155.

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

Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. l. Zypursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity Cell **101**:671-684.

Stuart, L. M., and R. A. Ezekowitz. 2008. Phagocytosis and comparative innate immunity: learning on the fly. Nat Rev Immunol **8**:131-141.

Vizoso, D. B., S. Lass, and D. Ebert. 2005. Different mechanisms of transmission of teh microsporidium Octosporea bayeri:a cocktail of solutions for the problem of parasite permanence Parasitology **130**:501-509.

Vlisidou, I., A. J. Dowling, I. R. Evans, N. Waterfield, R. H. ffrench-Constant, and W. Wood. 2009. Drosophila embryos as model systems for

monitoring bacterial infection in real time. PLoS Pathog **5**:e1000518.

Watson, L. F., F. T. Püttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel, and D. Schmucker. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects Science **309**:1874-1878

Yu, H. H., J. S. Yang, J. Wang, Y. Huang, and T. Lee. 2009. Endodomain diversity in the Drosophila Dscam and its roles in neuronal morphogenesis. J Neurosci **29**:1904-1914.

SUPPLEMENTARY MATERIAL

Region	Probe	Forward primer	Reverse primer
exon5	PROBE.EXON5.2 ATATTCGGGATGTTTCGCCGGAAG	EXON5.F CAAGTACATGGTTCTTCCCAGT	Ex5.2.R GGTCTCGCCAGTTAGACGAT
4.1		Ex4.1.2.F TCTCTTCAACATCCGACTGG	Ex5.1.R GTCCGGCATCGATAAGATTT
4.3		Ex4.3.2.F CCAAGTTGAATCGTGGATGA	
4.4		Ex4.4.1.F ACGACAATCACGTCGTTCAT	
4.2		Ex4.2.1F ACGGAACCGTCATTAACCAT	Ex5.2.R GGTCTCGCCAGTTAGACGAT
4.5		Ex4.5.3.F CGCAAATCTCGATACCCAGT	
4.6		Ex4.6.1.F ACTTACCACCCAACCGACAC	
4.8		Ex4.8.1.F TTTGTCATCGTCGACACTTG	
ß-actin	PROBE.ACTIN1 CCGTGAGAAGATGACCCAGATTATG	QUANT.ACTIN.F CGAGGAACATCCCGTTCTA	QUANT.ACTIN.1.R GTAGCCATCCAAGCAGTGC

Table S1 Primers and probes used in quantitative PCR (orientation 3' 5').

CHAPTER 3

POPULATION GENETICS OF DUPLICATED ALTERNATIVELY SPLICED EXONS OF THE *DSCAM* GENE IN *DAPHNIA* AND *DROSOPHILA*

Daniela Brites, Francisco Encinas-Viso, Dieter Ebert D, Louis Du Pasquier and Christoph Haag (2011). PLoS ONE 6(12): e27947. doi:10.1371/journal.pone.0027947

ABSTRACT

In insects and crustaceans, the Down syndrome cell adhesion molecule (Dscam) occurs in many different isoforms. These are produced by mutually exclusive alternative splicing of dozens of tandem duplicated exons coding for parts or whole immunoglobulin (Ig) domains of the Dscam protein. This diversity plays a role in the development of the nervous system and also in the immune system. Structural analysis of the protein suggested candidate epitopes where binding to pathogens could occur. These epitopes are coded by regions of the duplicated exons and are therefore diverse within individuals. Here we apply molecular population genetics and molecular evolution analyses using *Daphnia magna* and several *Drosophila* species to investigate the potential role of natural selection in the divergence between orthologs of these duplicated exons among species, as well as between paralogous exons within species. We found no evidence for a role of positive selection in the divergence of these paralogous exons. However, the power of this test was low, and the fact that no signs of gene conversion between paralogous exons were found suggests that paralog diversity may nonetheless be maintained by selection. The analysis of orthologous exons in *Drosophila* and in *Daphnia*, revealed an excess of non-synonymous polymorphisms in the epitopes putatively involved in pathogen binding. This may be a sign of balancing selection. Indeed, in *Dr. melanogaster* the same derived non-synonymous alleles segregate in several populations around the world. Yet other hallmarks of balancing selection were not found. Hence, we cannot rule out that the excess of non-synonymous polymorphisms is caused by segregating, slightly deleterious alleles, thus potentially indicating reduced selective constraints in the putative pathogen binding epitopes of Dscam.

INTRODUCTION

The gene encoding Down syndrome cell adhesion molecules (Dscam) has been studied in several metazoans. It codes for an integral membrane protein with signaling capacity, the extracellular part of which is formed by immunoglobulin (Ig) and fibronectin III (FNIII) domains. In insects and crustaceans *Dscam* evolved dozens of internal exon duplications which occur in three arrays (named arrays 4, 6, and 11 in *Daphnia* and 4, 6 and 9 in *Drosophila*) [1,2,3]. Due to a process of mutually exclusive alternative splicing, only one exon from each array is present in each mRNA molecule. This generates thousands of mRNA molecules coding for protein isoforms that differ in half of Ig2 (coded by any exon of array 4), half of Ig3 (coded by any exon of array 6), and in all of Ig7 (coded by any exon of array 11), while keeping the remaining domains constant (Fig. 1).

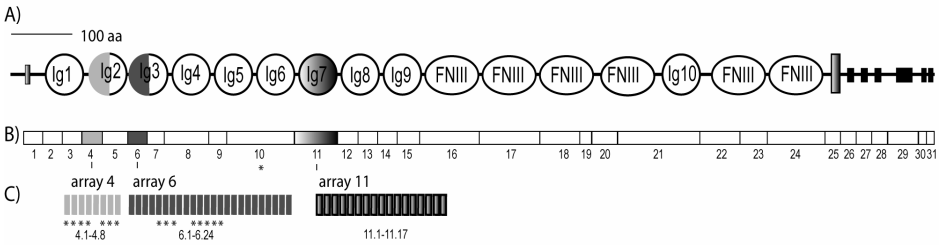


Figure 1 Dscam of *Daphnia* magna**.** A) Protein domains; Ig-immunoglobulin domains; FNIII- fibronectin III domains. The grey and black boxes represent the transmembrane and cytoplasmic domains. B) mRNA, each box corresponds to a constitutive exons and the colored boxes 4,6 and 11, correspond to exons that are the result of mutual exclusive alternative splicing of arrays of duplicated exons which are present in three arrays, as indicated in C) * Exons sampled in the present study.

In insects and crustaceans, the Dscam protein is believed to have a dual function acting both in the nervous system and in the immune system [1,2,3,4]. Its involvement in the nervous system development is well established in *Drosophila* where the different protein isoforms are essential for correct axon wiring [5,6]. The alternative splicing mechanism might be equally important for the immune function of Dscam: a diverse repertoire of Dscam isoforms is expressed in hemocytes, the immune cells of insects and crustaceans, and these isoforms can bind different bacteria depending on exon composition [1,7]. Furthermore, the splicing patterns of the alternative exons change upon infection, and silencing of Dscam leads to lower phagocytosis rates in *Drosophila* and *Anopheles* [1,4]. However, Dscam does not seem to be required

for *E. coli* phagocytosis in *Drosophila* embryos [8]. Given that the hemocytes of adult flies are of embryonic origin these results are somewhat controversial. On the other hand, the partial blockage of bacteria uptake [1] suggests that phagocytosis is not under the control of a single pathway and it is possible that DSCAM-silenced individuals [1] behave differently from *dscam05518* mutant embryos [8] where a surrogate mechanism may take over.

The first four Ig domains of the Dscam protein form a stable horse-shoe structure, which is probably common to all isoforms [9], Fig. 2a). Parts of Ig2 and Ig3 together form two surface epitopes at either side of the horse-shoe structure, epitope I and epitope II. Both epitopes are partly coded by array 4 and partly by array 6 (Fig. 2b, Fig. S1). Epitope I is crucial for the formation of Dscam dimers and for the development of the nervous system [9]. Epitope II is oriented towards the external environment of the Dscam molecule, and is thus a candidate epitope for the interaction with antigens.

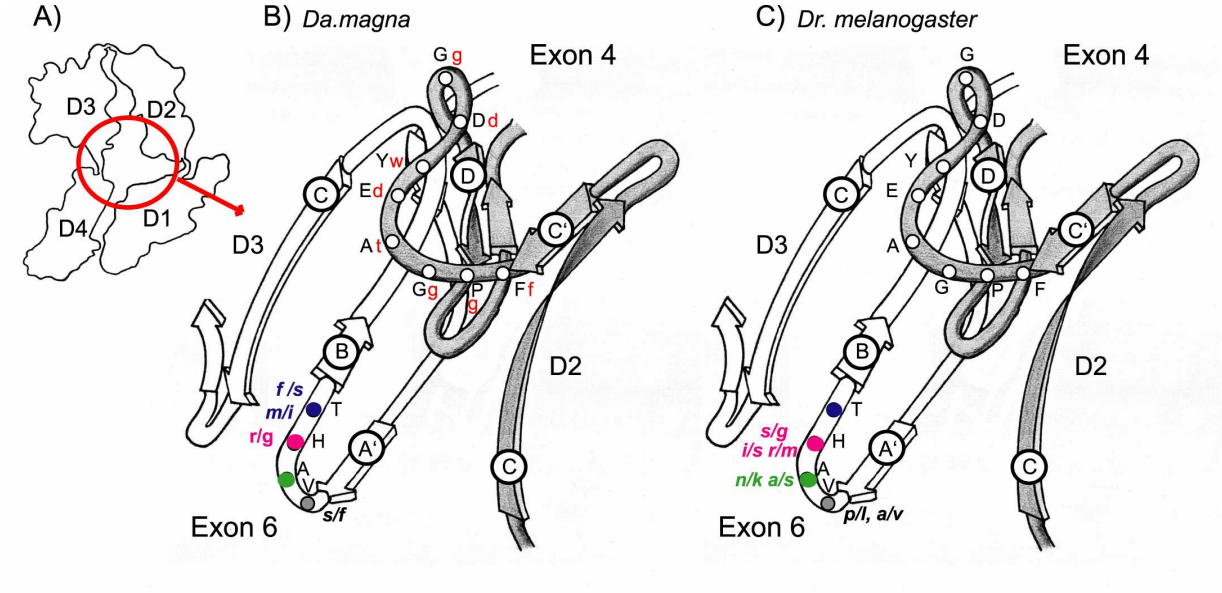


Figure 2 A) Outline of the Dscam horse-shoe structure formed by the first four Ig domains (D1-D4). B & C) Detail of Epitope II, formed by the two interstrand loops C'-D of exon 4 and A'-B of exon 6, respectively. Each strand is indicated by an encircled letter. The *Drosophila* aminoacid residues corresponding to the actual structures are in black uppercase initials (exon 4.1 and 6.34 of *Dr. melanogaster*). *Da. magna* residues have been positioned in function of the known homology of the molecule in the region coded by exon 4 and 6 (BRITES *et al.* 2008) and are represented by red lowercase initials. Polymorphic sites at exons 6 for *Da. magna* and *Dr. melanogaster* are represented by lowercase initials, each color corresponds to positions on Epitope II coding regions in different paralogous exons 6.

The sequence of each exon belonging to arrays 4 and 6 can be divided into parts of the sequence that contribute to epitope I, parts that contribute to epitope II, and parts that contribute to neither of them. Orthologous exons of arrays 4 and 6 show more divergence between closely related *Drosophila* species in the parts coding for epitope II than in the parts coding for epitope I [9]. This pattern, in combination with the structural features described above, has led to the idea that epitope II might be involved in host-parasite coevolution and might have evolved faster as a consequence of being a potential pathogen recognition epitope [9]. Here we address this hypothesis by searching for signatures of adaptive evolution in the nucleotide sequence coding for epitope II. We do this by analyzing polymorphism patterns of the Dscam gene in *Daphnia magna* and *Drosophila melanogaster* as well as divergence patterns between these species and some of their closely related congeners and by using molecular tests of selection, including maximum likelihood (ML) models of codon evolution.

MATERIAL AND METHODS

Origin of the samples

We used 17 genotypes of *Da. magna*, each isolated from a different population, as well as one genotype from two outgroup species, *Da. lumholtzi* (Zimbabwe) and *Da. similis* (Israel) (Table 1). The genotypes were maintained by clonal propagation of offspring from single females isolated from these populations.

The polymorphism data for *Dr. melanogaster* were obtained by [10] and come from six populations (four individuals per population pooled before DNA extraction), covering the initial range of the species in Africa and more recent expansions. The divergence data for *Drosophila* are from the sequenced genomes of six species of the *melanogaster* group obtained from gene bank (*Dr. ananassae* GF12235; *Dr. melanogaster* CG17800; *Dr. erecta* GE24114; *Dr. simulans* FBgn0086259; *Dr. yacuba* GE24114; *Dr. sechellia* CH480816). *Daphnia pulex* and other *Drosophila* species were not considered for the analysis because their synonymous site divergence was too high to allow a meaningful analysis of substitution rates due to the high likelihood of multiple hits. However, the following six additional species were included in analyses of exon copy number and analyses based on amino acid sequences only (where multiple hits are much less likely than at synonymous sites): *Dr. pseudoobscura* (GA14672), *Dr. persimilis* (CH479181), *Dr. willistoni* (CH963849), *Dr. mojavensis* (GI20826), *Dr. virilis* (GJ20560), *Dr. grimshawi* (CH916367).

Genomic region analyzed

In *Da. magna* the entire Dscam protein, depending on exon usage, is composed of

approximately 1960 amino acids and the whole locus is 31 Kb long [3]. For the present study, we analyzed three regions of the Dscam gene: two regions containing alternatively spliced, duplicated exons belonging to arrays 4 and arrays 6 (and, for comparison, one region containing the constitutive exon 10, which was chosen because it codes for Ig6, which is structurally similar to the Igs 2 and 3, coded for by arrays 4 and 6 (data not shown).

In *Da. magna*, array 4 consists of eight paralogous exons, (named 4.1 to 4.8, covering around 3390 bp in total) and array 6 contains 24 paralogous exons (6.1 to 6.24, around 6100 bp in total). We obtained sequence data on all exons of array 4, except exon 4.5 (3200 bp in total, accession numbers JN977549 to JN977579)), exons 6.5 to 6.7 and 6.10 to 6.14 (1683 bp in total, accession numbers JQ037914 to JQ037973), and 327 bp of the constitutive exon 10 (the total length of which is 423 bp, accession numbers JQ037974 to JQ037993). Part of the intron sequences (mostly from array 4) had to be excluded from the analysis due to alignment ambiguities, repetitive sequences, and insertion/deletion polymorphisms. Thus, only 1759 bp of array 4 sequences and 1679 bp of array 6 sequences were retained for analysis (Table 2). All exons sampled are known to be expressed [3]. The same sequence data was also obtained for one genotype of *Da. lumholtzi*. We were unable to obtain array 6 sequence from *Da. similis*, thus we restrict the analysis of between-species divergence mostly to divergence between *Da. magna* and *Da. lumholtzi* which is the closest known species to *Da. magna*

Insects have three other Dscam paralogs that have been named Dscam-like (Dscam-L) [3,11,12] and we have found orthologues of these *Dscam-L* genes in the genome of *Daphnia pulex* (unpublished data). The distinction between the variable *Dscam* and the *Dscam-L* genes is very clear and we are confident that we have amplified only the variable Dscam in *Daphnia*.

The Dscam sequence data from *Dr. melanogaster* [10] comprises almost the entire Dscam coding region (22795 bp). For the interspecific comparisons of the six *Drosophila* species from the melanogaster group, we used all orthologous exons of arrays 4 (12 exons, 1950 bp in total). For array 6, 43 orthologous exons were used, 32 occurring in all six species and eleven in five of them (5205 bp in total). Exons that confidently (>60% of 100 bootstrap replicates) shared a common ancestor in a maximum likelihood tree were considered orthologous [13]. Trees were built with RAxML trough the Cipres Portal [14].

Sequencing methods

Genomic DNA of *Daphnia* genotypes was extracted (peqGOLD Tissue DNA Mini Kit, PEQLAB, Erlangen, Switzerland) and PCR reactions were carried out using High Fidelity Polymerase (ROCHE, Manheim, Germany) for array 4 exons or Pfu (PROMEGA, Madison, WI, USA) for array 6 exons and exon 10. Primers

and PCR conditions are available by request. PCR products were purified (Gen Elute^{TM} PCR Clean-up kit, SIGMA, St Louis, MO, USA), and all reactions were sequenced directly using Sanger sequencing. In addition, products of some PCR reactions were cloned (TOPO Kit, INVITROGEN, Carlsbad, CA, USA) to obtain experimental haplotype information. All heterozygous sites and singleton polymorphisms were confirmed by resequencing independent PCR reactions or cloning. To verify that only the targeted regions were amplified, all sequences were compared to a reference *Dscam* sequence, obtained by cloning the entire locus in *Da. magna* [3]. The *Dscam* sequence data from *Dr. melanogaster* was obtained by Solexa-Illumina sequencing [10]. Regions with less than 20x coverage were excluded. By resequencing eleven genes using Sanger sequencing, the authors uncovered 31 miscalled polymorphic sites in a total of 12451 bp (accuracy=99.8%), of which 10 polymorphisms (0.08%) corresponded to false positive polymorphisms and the remaining to false negatives (0.12%) [10]. To minimize the occurrence of false positives all variants with a frequency of less than 5% within a population were excluded from the analysis [10]. Because read frequencies did not provide a reliable estimate of allele frequencies [10], the data were only used to estimate nucleotide diversity from the proportion of segregating sites (θ) and for performing McDonald-Kreitman tests [36], but not for tests based on allele frequencies.

Identification of epitope I and epitope II coding sequences

Some analyses required partition of array 4 and array 6 exon sequences in regions that constitute epitope I, epitope II, and the remaining exon regions. These partitions were based on the structural information provided by [9] and on the similarities in the secondary structure of Dscam between *Da. magna* and *Drosophila melanogaster* (data not shown), using the program PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) [15]. The partitions were assigned in the following way: In exons of array 4, the ten amino acids between the conserved 4Q and the 15V were considered to belong to epitope I, and the 13 amino acids after 40W were considered to belong to epitope II. In exons of array 6, the eight amino acids after 10R were considered to belong to epitope I, and the eight amino acids before the conserved LLC motive were considered to belong to epitope II (Fig. S1). Figure 2 was redrawn manually from [9] using the Dscam reference (2v5m) in the protein data bank (PDB, http://www.rcsb.org/pdb/home/home.do).

Analysis

Sequences were assembled and edited using STADEN version 1.5 (http://staden.sourceforge.net/), aligned with ClustalX [16] and edited in Jalview 2.3 [17]. For exons of array 6, alignments including unphased sequences (7 genotypes) and true haplotypes (20 cloned haplotypes) were used to obtain

pseudohaplotypes for unphased sequences using the program PHASE 2.1 [18]. For array 4 exons all PCR products were cloned. The program GENECONV version 1.81a (using default parameters) was used to detect gene conversion between paralogous exons [19].

Analyses of nucleotide diversity (π), divergence, and standard neutrality tests were done with DNAsp v5 [20]. Unless stated otherwise, divergence always refers to divergence of orthologous sequence between species, rather than divergence of paralogous sequence within species. Amino acid divergence between paralogous exons was calculated using the Poisson correction method to account for multiple substitutions at the same site, averaging over all paralogous pairs MEGA 4.0 [21]. Next, we used the site models implemented in PAML version 4 [22,23] and HYPHY [24,25] to test for positive selection between orthologous exons using six *Drosophila* species from the *melanogaster* group. The same models were not applied to *Da. magna* because they require data from several, closely related species. These methods assess the ratio of non-synonymous to synonymous substitutions $\omega = dN/dS$, where $\omega < 1$ indicates purifying selection, $\omega = 1$ neutrality, and $\omega > 1$ positive selection. They infer positive selection by asking whether a model that allows some codons to have $\omega > 1$ fits the data significantly better than a model that restricts all codons to have $\omega \leq 1$.

The ML analysis was carried out in the following way: In PAML, we calculated likelihoods for the following models: M1a (assuming that sites have either $0<\omega o<1$ or $\omega 1=1$), M2a (which adds an additional class of sites with $\omega 2 > 1$), M7 (which uses a ß-distribution to model ω and does not allow for $\omega>1$), and M8 (which adds an extra class of sites with $\omega>1$ to M7). We compared the log-likelihoods between models M2a and M1a and between M8 and M7 to test for positive selection [23]. In all models, base frequencies were calculated from the average nucleotide frequencies at the three codon positions and we used the GY model [26] as basic model of codon substitution. Finally, we used the empirical Bayes approach implemented in PAML to identify individual codons under positive selection.

To account for potential differences in synonymous rates, which can influence the accuracy of detecting positively selected sites, we fitted the "dual" model implemented in HYPHY to our data [25]. We used a general discrete distribution (GDD) with three bins for dN and dS and the codon substitution model MG94 [26] combined with the nucleotide substitution model HKY85 (determined as the best-fitting nucleotide substitution model using the model selection procedure implemented in HYPHY). To identify sites under selection we used a Bayes factor of 50.

To test whether the dN/dS of epitope II regions differed from remaining of exon regions (for a similar analysis see [27] [28], we applied the ML-based hypothesis testing procedure

implemented in HYPHY on two partitions of the data, one containing epitope II sequence and one containing the remaining sequence of the exons. The same tree topology and the MG94 codon model combined with HK85 nucleotide substitution model were assigned to each partition (epitope II and non-epitope II sequence) considering the observed nucleotide frequencies. For testing the hypothesis that dN/dS differs between partitions, dN/dS was estimated independently for each of them but the same tree was assumed.

To investigate substitutions patterns of paralogous exons, we applied branch models [29,30] as implemented in PAML. This analysis was performed only on the phylogeny of exons of array 6 in the *Dr. melanogaster* group (Fig. S3 A). Paralogous exons 4 have diverged too much for a reliable analysis (data not shown).Whereas orthologous exons 6 are very conserved (except epitope II coding regions), paralogous exons diverged extensively pointing out to an acceleration of aminoacid substitutions following exon duplication. Using the branch models on trees that included orthologous as well as paralogous sequences, allowed us to test whether selection changed after duplication by contrasting branches giving rise to paralogs with branches giving rise to orthologs. We used an alternative model assuming that orthologous branches and paralogous branches differ in ω (model R2, Fig. S3 A & B), the null hypotheses being that all branches in the tree have the same ω (model R1, Fig. S3 A & B). Under these models, ω estimates correspond to an average over branches and sites and thus unlikely to be higher than 1. We used the branch-site models implemented in PAML to test for positive selection, i.e. to test whether particular branches have aminoacid sites that evolved with a $\omega>1$ [31,32]. Because we did not have *a priori* data on particular exons with functional importance we chose to test the branches leading to duplicated exons where we detected an excess of non-synonymous polymorphism in *Dr. melanogaster* using MK-tests in the previous analysis. For doing this, smaller subtrees were used (Fig. S3 A).

TABLE 1 Geographic origin of the *Da. magna* populations sampled

Genotype	Geographic origin	Latitude	Longitude
FA	Tvärminne, Finland	59°50.18'N	23°14.16'E
K-10-1	Tvärminne, Finland	59°49.43'N	23°15.15'E
SP1-2-3	Tvärminne, Finland	59°48.42'N	23°12.31'E
FAV-1-1[1]	Åland Islands, Finland	60°01.30'N	19°54.15'E
HO1[1]	Hungary	46°48'N	19°08'E
HO2	Hungary	46°48'N	19°08'E
HO3[1]	Hungary	46°48'N	19°08'E
DKN-1-8	Kniphagen, Germany	54°10.45'N	10°47.3'E
MU10	Munich, Germany	48°12.23'N	11°42.34'E
MU11	Munich, Germany	48°12.23'N	11°42.34'E
GE-1	Ismaning, Germany	48°12.23'N	11°42.34'E
SC1	Leitholm, UK	55°43.9'N	02°20.43'W
EC-1-4	Cummor, UK	51°43.9'N	01°20.4'W
CN-2-1	Sedlec, Czech Republic	48°46.52'N	16°43.41'E
BE-OM-1	Leuven, Belgium	50°52'N	04°41'E
KE-1	Kenia	0°26.25'N	35°18.16'E
SE-2-3	Sweden, East coast	60°25.93'N	18°31.34'E

[1] Genotypes for which only array 6 exons were amplified, and which were only used in parts of the analysis.

TABLE 2 Number of sites and number of polymorphic sites per Dscam genomic region analyzed in *Da. magna* **(Dmag)** and *Dr. melanogaster* (**Dmel**), the latter obtained from [10]

Gene region	N of sites (L)					N of polymorphic sites (S)				
	Dmag			**Dmel**		**Dmag**			**Dmel**	
	L_s	L_a	L_{nc}	L_s	L_a	S_s	S_a	S_{nc}	S_s	S_a
Array 4 total	218	731	778	458	1524	4	6	20	11	9
Epitopes I	34	117	n.a.	n.a.	n.a.	0	0	n.a.	n.a.	n.a.
Epitopes II	56	187	n.a.	120	447	2	1	n.a.	2	4
Remaining	128	427	n.a.	338	1077	2	5	n.a.	9	5
Array 6 total	213	628	728	1443	4325	17	10	27	60	46
Epitopes I	44	124	n.a.	n.a.	n.a.	1	1	n.a.	n.a.	n.a.
Epitopes II	40	128	n.a.	278	864	0	5	n.a.	29	17
Remaining	129	376	na	1164	3461	16	4	na	77	29
Ig6 coding exon	81	246	0	60	173	6	4	0	25	0

Abbreviations: n.a., not assessed; $_s$, synonymous; $_a$, non-synonymous; $_{nc}$, non-coding.

RESULTS

Gene conversion and copy number of array 4 and array 6 exons

The duplicated exons of are 160 bp in array 4 and 130 bp in array 6, and within each array, they are separated by introns of approximately 200 bp (array 4) and 100 bp (array 6)**.** None of our PCRs showed evidence (length polymorphism or failed PCRs) for variation in the number of exons in array 4, nor in array 6 (only eight contiguous exons out of 24 were investigated in the latter). We found no variation among closely related species in the number of paralogous exons in array 4: all twelve *Drosophila* species have twelve exons whereas both *Da. magna* (EU307883) and *Da. pulex* (EU307884) have eight. In contrast, array 6 has between 41 and 52 exons in the twelve *Drosophila*, and two more exons in *Da. pulex* than in *Da. magn*a. Furthermore, in *Da. lumholtzi*, at least one of the eight sampled exons of array 6 is probably missing (as indicated by our failure to obtain this sequence). This

indicates that exon copy number in array 6, but not in array 4, varies among related species.

Multigene families are frequently under the action of concerted evolution by gene conversion [33]. However, consistent with earlier results based on trees of the duplicated regions in *Da. magna* and *Da. pulex* [3], we found no evidence for gene conversion between duplicated exons in arrays 4 and 6 (*p*-values based on 10000 permutations were 0.2 for array 4 and 0.5 for array 6). The low levels of polymorphism in array 4 (Table 3) may suggest gene conversion, but the high level of divergence between paralogous exons (Table 3) contradicts this hypothesis. The apparent absence of gene conversion suggests that Dscam is unusual in this respect compared with other multi-gene families and greatly facilitates further analysis because it legitimates the use of classical population genetic methods.

General patterns of polymorphism and divergence

In *Da. magna*, array 4 has low nucleotide diversity (π) both at non-synonymous and at synonymous sites, whereas array 6 and exon 10 have moderate levels of synonymous diversity (π_s) (Table 3), similar to the average values estimated for eight housekeeping *Da. magna* genes in another study [34], and higher than in a sample of putative immunity genes in this species [35]. In contrast, non-synonymous diversity (π_a) in array 6 and exon 10 is about ten times higher than in other *Da. magna* genes [34]. Synonymous divergence (k_s) between *Da. magna* and *Da. lumholtzi* is similar in all sampled Dscam regions. Contrastingly, non-synonymous divergence (k_a) is much higher in arrays 4 and 6 than in exon 10, and correspondingly also ka/ks ratios are higher in arrays 4 and 6 than in exon 10 (Table 3). The opposite is true for the ratio of non-synonymous to synonymous nucleotide diversity ratio ($\pi_{a/}\ \pi_s$, Table 3).

TABLE 3 Estimates of Dscam nucleotide diversity (π in *Da magna*, θ in *Dr melanogaster*), divergence of orthologous sequences between *Da. magna* and *Da. lumholtzi*, and amino acid divergence between paralogous regions of *Da. magna*, as well as divergence of orthologous sequences between *Dr. melanogaster* and a reconstructed ancestral sequence estimated in [10].

Species	Gene region	Diversity (π, θ)				Divergence (k) [2]				
	Array 4 Total	0.0014	0.004	0.005	0.0008	0.2	0.132	0.013	0.098	0.837
	Epitopes I	0	n.a.	0	0	n.a.	0.118	0.000	0	0.980
	Epitopes II	0.0014	n.a.	0.005	0.0009	0.18	0.164	0.032	0.195	1.431
	Remaining	0.0014	n.a.	0.005	0.0004	0.08	0.137	0.004	0.029	0.567
Dmag	Array6 Total	0.0064	0.01	0.017	0.003	0.176	0.148	0.013	0.088	0.593
	Epitopes I	0.003	n.a	0.003	0.0006	0.1	0.139	0.008	0.057	1.379
	Epitopes II	0.007	n.a.	0.000	0.009	n.a.	0.178	0.031	0.174	1.616
	Remaining	0.007	n.a.	0.023	0.001	0.04	0.144	0.004	0.028	0.211
	Exon10 (Ig6)	0.006	n.a.	0.011	0.005	0.454	0.149	0.003	0.02	n.a.
	Array 4 Total	0.01	n.a.	0.024	0.006	0.25	0.039	0.003	0.077	n.a.
	Epitopes II	0.0106	n.a.	0.017	0.009	0.53	0.033	0.005	0.151	n.a.
	Array 6 Total	0.018	n.a.	0.042	0.011	0.26	0.076	0.008	0.105	n.a.
	Epitopes II	0.0253	n.a.	0.043	0.006	0.14	0.082	0.01	0.121	n.a.
Dmel [6]	Exon7 (Ig6)	0.008	n.a.	0.033	0	n.a.	0.083	0	n.a.	n.a.
	Remaining Dscam[4]	0.019	n.a.	0.048	0.009	0.18	0.067	0.005	0.075	n.a.
	Control genes [5]	n.a.	n.a.	0.015	0.002	0.13	n.a.	n.a.	n.a.	n.a.
	Immune genes [5]	n.a.	n.a.	0.016	0.009	0.56	n.a.	n.a.	n.a.	n.a.

Abbreviations: n.a., not assessed; $_{;t}$ total; $_{s}$ synonymous; $_{a}$ non-synonymous; $_{nc}$ non-coding
[1][34], average over eight housekeeping genes; [2] Divergence estimates are not corrected for diversity within species nor for multiple hits; [3]amino acid divergence between paralogous regions of *Da. magna*. [4]from Ig2 coding exons to the first transmembrane domain coding exon, except arrays 4 and 6 coding exons (total of 15045bp). [5]estimates by [10]; [6] Data obtained by [10].

The divergence estimates between *Da. magna* and the second outgroup species, *Da similis* are similar to the estimates between *Da. magna* and *Da. lumholtzi*. Thus they are presented in the supplementary materials only (Table S5) and will not be discussed further. A McDonald and Kreitman (MK)-test [36] yielded evidence for an excess of non-synonymous polymorphism compared to the ratio between non-synonymous and synonymous divergence in array 4, whereas results for array 6 and exon 10 did not differ from neutral expectations (Table 4). This is consistent with the action of balancing selection in array 4, but a Hudson-Kreitman-Aguadé (HKA) test [37] did not yield evidence for a significantly higher polymorphism to divergence ratio in array 4 compared to array 6 and exon 10 combined (synonymous sites only, p=0.08). All non-synonymous polymorphisms in array 4 segregate at low frequencies (Table S1), so that the excess of non-synonymous polymorphism could also reflect slightly deleterious mutations. In such cases it has been suggested that removing. alleles with a frequency lower than 0.15 from the MK analysis could partially reduced the bias introduced by low-frequency polymorphisms [38]. When applying this to our data, only exon 10 has a significant excess of non-synonymous polymorphism.

In *Dr. melanogaster*, non-synonymous diversity is similar to that of other genes with immunity-related functions, and synonymous diversity is higher than that of other immune and control genes [10] (Table 3). In contrast to *Da. magna*, constitutively expressed and alternatively spliced exons exhibited similar levels of synonymous and non-synonymous diversity. A MK-test applied to arrays of exons 4 and 6 revealed an excess of non-synonymous polymorphism in relation to what would be expected from the divergence levels between *Dr. melanogaster* and an inferred ancestral sequence [10]. After eliminating all alleles that occurred with minor frequencies (less than 0.15) there was no longer an indication of a significant excess of non-synonymous polymorphisms in relation to divergence (Table 5).

Contrasting patterns in Epitopes I and II

In *Da. magna* non-synonymous polymorphism was higher in epitope II than in the other regions (Table 3). Likewise non-synonymous divergence is nearly an order of magnitude higher in epitope II compared to epitope I and the remaining exon regions and also compared to exon 10 (Table 3). Contrastingly, synonymous site divergence between *Da. magna* and *Da. lumholtzi* was similar for epitope I, epitope II, and the remaining exon regions of arrays 4 and 6 (Table 3). However, neither the MK-test on epitope II nor the HKA-test comparing epitope II to all remaining regions indicated a significant deviation from neutrality, although there was a tendency for excess non-synonymous polymorphism in epitope II (Table 4). When array 6 was considered alone, this excess of non-synonymous polymorphism was significant

(p=0.04, Table 4), mostly due to exon 6.7 (Fig. S2). This effect disappeared, however, if alleles with a frequency lower than 0.15 were excluded from the analysis (Table 4).

Likewise, in *Dr. melanogaster* array 6 epitope II coding regions exhibited a significant excess of non-synonymous polymorphism relative to the levels of divergence estimated between *Dr. melanogaster* and an inferred ancestral sequence [10]. After removing minor allele frequencies (less than 0.15), the excess of nonsynonymous polymorphism was stronger because mainly synonymous mutations were excluded (Table 5). It is not possible to accurately estimate allele frequencies from the data obtained by [10] in order to know whether the non-synonymous derived alleles are common in the populations analyzed. However, the same derived non-synonymous alleles are present in several of the *Dr. melanogaster* populations surveyed around the world suggesting that they are not rare variants (Table S3).

TABLE 4 MacDonald Kreitman tests for the comparison between *Da. magna* and *Da. lumholtzi*. The test was performed on raw frequencies of alleles as well on frequencies after correcting for minor allele frequency (MAF). This correction was done by eliminating all allele frequencies lower than 0.15 when considering all *Da. magna* populations.

Gene region	Raw values					Corrected MAF				
	Fixed		Polymorphic		p^1	Fixed		Polymorphic		p^1
	Syn	Nonsyn	Syn	Nonsyn		Syn	Nonsyn	Syn	Nonsyn	
Array 4 Total	28	9	4	6	0.05	28	9	1	0	1
Epitopes II	10	7	2	2	1	10	7	0	0	n.a.
Array 6 Total	26	7	17	10	0.25	29	7	4	2	0.6
Epitopes II	6	4	0	5	0.04	6	4	0	2	0.4
Exon 10 (Ig6)	10	0	6	4	0.08	12	0	0	2	0.01

[1] *p* values are according to a two-tailed Fisher's exact test. n.a., not assessed.

Testing for positive selection in epitope II regions in *Drosophila*

The ML analysis implemented in PAML and HYPHY did not yield significant evidence for positive selection in arrays 4 and 6 in the *melanogaster* group, when the entire orthologous coding regions of the two arrays were analyzed, (Table 6, HYPHY results not shown). When the dN/dS of epitope II coding regions was contrasted with the remaining exon regions for both arrays of exons 4 and 6 (Table 6), a model that estimated dN/dS separately for epitope II and for the remaining regions fitted the data better than a model that considered dN/dS to be constant throughout the entire exons. The dN/dS estimates of epitope II coding regions were significantly higher than for the remaing regions,

but not higher than 1 (p<0.001 in both cases, Table 6).

Divergence between paralogues

The selective constrains acting before and after the duplications of exons 6 differed according to our branch model analysis (Table S4, p<0.001). The average ω over all sites and branches leading to paralogous exons was 0.26 whereas the branches leading to orthologous exons had average ω of 0.094. The branch site analysis on several branches did not provide evidence for a role of positive selection in the divergence between the paralogues (Table S4).

TABLE 5 MacDonald Kreitman tests for the comparison between *Dr. melanogaster* and an ancestral sequence inferred by [10]. The test was performed on raw frequencies of alleles as well on frequencies corrected for minor allele frequency effects (MAF). This correction was done by eliminating all allele frequencies lower than 0.15 when considering all *Dr. melanogaster* populations.

Gene region	Raw values				*P*	Corrected MAF				p^1
	Fixed		Polymorphic			Fixed		Polymorphic		
	Syn	Nonsyn	Syn	Nonsyn		Syn	Nonsyn	Syn	Nonsyn	
Array 4 Total	13	0	11	9	0.005	13	0	5	0	n.a
Epitopes II	3	0	2	4	0.16	3	0	0	0	n.a
Array 6 Total	81	14	60	46	<0.001	86	18	18	8	0.1
Epitopes II	17	7	12	17	0.051	19	7	2	7	0.01
Exon 7 (Ig6)	4	0	2	5	n.a	4	0	1	0	n.a

[1] *p* values are according to a two-tailed Fisher's exact test. n.a., not assessed

TABLE 6 Likelihood ratio tests and maximum likelihood estimates of dN/dS for six *Drosophila* species of the *melanogaster* group.

Gene region (Models tested)	N° variable sites	LRT	Parameter estimates
Array 4 total			
(M1a[1] vs. M2a[2])	292	n.s.	ω_0=0.009 (96%) [3]
(M7 vs. M8)			$\omega_{1\&2}$=1 (4%) [3]
Epitopes II	84	χ2=52 [4];df=1;	dN/dS=0.11
Remaining	208	p<0.001	dN/dS=0.006
Array 6 total			
(M1a[1] vs. M2a[2])	784	n.s.	ω_0=0.03 (94%) 3
(M7 vs. M8)			$\omega_{1\&2}$=1 (6%) 3
Epitopes II	242	χ2=119 [4];df=1;	dN/dS=0.19
Remaining	542	p<0.001	dN/dS=0.03

Abbreviation: LRT, Likelihood ratio tesrt

[1] M1a: ω_0 varies between 0 and 1 whereas ω_1=1; [2] M2a adds to M1a, ω_2>1, which is estimated from the data; [3] proportions of sites under ω_0, ω_1, and ω_2. [4] Tests whether the dN/dS relative to the two partitions are significantly different from each other.

DISCUSSION

Insights into exons duplications in arrays 4 and 6

The duplicated exons of arrays 4 and 6 contribute to Dscam isoform diversity due to alternative splicing [11]. Selection on duplicated genes occurs at two levels: on copy numbers and on new mutations within the duplicated forms [39]. In *Daphnia*, we did not find any copy number polymorphism in array 4 among closely related species. This is consistent with results from insects, which indicate that the structure of array 4 is ancient and remained relatively unchanged throughout the evolutionary history of insects [40]. In contrast, the number of exons in array 6 is larger than in array 4 [40] (this study). The reasons for these differences are unknown and our results do not allow distinguishing whether constraints or adaptive evolution might explain them.

Much of the sequence diversification of paralogous exons in arrays 4 and 6 seems to have predated the most recent speciation events, and, in both arrays, exons do not seem to have undergone much concerted evolution, but rather evolved under a birth-and-death evolution process [3]. This is supported by the apparent absence of recent gene conversion events, which is surprising as gene conversion occurs in the majority of other multi-copy gene families [33]. Likely there is selection against gene conversion because it would homogenize exon sequences, thus diminishing the repertoire of different Dscam isoforms. Functional studies showed that Dscam isoform diversity is indeed necessary for the correct development of the nervous system [5]. Interestingly, other important multi-copy

immunity related gene families, such as MHC, immunoglobulins, and T-cell receptors, evolve also mainly by birth-and-death evolution rather than by concerted evolution [33].

Polymorphism and divergence in arrays 4 and 6

Standard tests did not provide evidence for positive selection in arrays 4 and 6 as a whole in *Da. magna*. Rather, all three studied regions showed a tendency for an excess of non-synonymous polymorphism (significant only for array 4). While this can be interpreted as an indication of balancing selection, most of the non-synonymous polymorphisms segregate at low frequency, so that they may also represent segregating, slightly deleterious variants [38]. Also in *Dr. melanogaster*, the excess of non-synonymous polymorphisms in arrays 4 and 6 is mainly caused by low frequency variants. This might derive from the action of purifying selection on the alternatively spliced exons being weaker than on constitutively expressed exons because the former are less expressed than the latter. Yet, rare alleles may also be maintained by time-delayed negative frequency dependent selection which has been described for host-parasite systems [41, 42]. Under this kind of selection, there is a time lag between the allele frequencies and the selection acting on the allele, so that (in contrast to e.g., overdominant selection), allele frequencies are expected to fluctuate in different populations and alleles can be rare for a considerable amount of time [41, 42]. Furtermore, sporadic fixation of alleles may occur and low synonymous variation is predicted due to bottlenecks for the different alleles [43]. Consistent with this prediction, in *Da. magna*, array 4 exons have low synonymous variation. However, in contrast *Dr. melanogaster* tends to have high synonymous variation across the entire *Dscam* gene (Tab. 3).

The evolution of epitopes I and II

Structural data suggest that epitope I is a crucial unit engaged in the formation of Dscam homologous dimers between the surface of neurons, whereas epitope II is oriented towards the outside of the Dscam protein and is a putative antigen binding region [9]. Within species, the paralogous exon regions of arrays 4 and 6 coding for epitopes I and II have diverged more than the remaining regions of the gene (Table 3). In contrast, divergence between orthologous exon regions coding for epitopes I is much lower than between orthologous exon regions coding for epitopes II in both *Daphnia* (this study) and *Drosophila* [9]. These patterns suggest that the divergence between paralogs is ancient. Intriguingly, however, epitopes I do not seem to have evolved much since then, except by exon duplications, whereas epitopes II have continued to accumulate differences, which is seen in the increased divergence of orthologous sequence between closely related species (Table 3).

Potential balancing selection in epitopes II

While much of the sequence divergence between paralogous exons may be ancient, allowing high isoform diversity, divergence driven by selection may still be ongoing in some parts of the gene, particularly if any parts of the gene are involved in ongoing coevolution with parasites. Epitope II coding regions of exons 6 in both *Daphnia* and *Drosophila*, show an excess of nonsynonymous polymorphisms relative to the divergence levels. In *Dr. melanogaster*, this effect is still visible after excluding low frequency alleles and may thus suggest balancing selection [44]. In *Dr. melanogaster* allele frequencies could not be inferred with great accuracy, but we found that the same derived non-synonymous alelles segregate in the several *Dr. melanogaster* populations around the world, which suggests that these alleles are not slightly deleterious and are not artifacts due to PCR or sequencing errors (Table S3). Additionally, some of these alleles are present in other distantly related *Drosophila* species, raising the possibility that some of those could be trans-specific polymorphisms (Table S3). However, we did not find high levels of non-synonymous nucleotide polymorphism in Epitope II coding regions, in contrast to that found in the resistance genes *APL1* and *TEP1* of *Anopheles gambiae* to *Plasmodium falciparum*, whose very high levels of non-synonymous polymorphism are presumably a result of balancing selection and gene conversion [45,46].

If balancing selection is maintained for a long time, it is expected to lead to strong linkage disequilibrium (LD) and to elevated neutral variation at linked sites [44,47]. In *Da. magna* the synonymous site diversity of exon 6.7 is among the highest of all sampled exons in array 6 (π_s = 0.012), but synonymous site diversity of the whole array 6 is only slightly higher than that of the constitutive exon 10. In addition, we did not find elevated LD in the region (results not shown). Thus if any balancing selection acts on the region, it is unlikely to be long-term balancing selection, as found in some other immunity genes such as MHC [48]. In the *Dr. melanogaster* populations, Dscam synonymous diversity tends to be high across the whole gene (Table S2), but it is not possible to estimate whether there are any sites in LD with epitope II coding sites given that no haplotype information is available.

An alternative explanation, as discussed above, is that epitopes II are under negative frequency dependent selection. In such case, due to periodic bottlenecks, non-synonymous diversity is not expected to be elevated [43] and the prediction for LD is less clear. However, to differentiate between overdominant and negative frequency dependent selection acting on this region would require better estimates of allele frequencies among different populations both in *Daphnia* and *Drosophila*. In summary, our data do not currently allow us to distinguish between the hypothesis of negative frequency-dependent selection and the hypothesis of relaxed selective constraints, although the fact that the same derived alleles segregate in several *Drosophila*

populations suggest a likely action of some form of balancing selection.

Maximum likelihood codon based site models have been shown to be powerful at detecting balancing selection in MHC [28,49]. Yet many of the studies on MHC involved comparison of paralogous MHC alleles [48,50] [28,49]. In Dscam, paralogous exons diverged too extensively (array of exons 6 tree length for dS is 104.4 in *Dr. melanogaster*) to be included in a reliable site model analysis [51]. The site model analysis of orthologous exons of arrays 4 and 6 in six *Drosophila* species revealed that although epitopes II evolve faster than the remaining regions of these arrays, there is no evidence that this is driven by positive selection. However, as discussed in the supplementary section (Table S2), our analysis has most likely low power for detecting balancing selection.

Involvement of epitope II in immune recognition in insects and crustaceans

Despite some differences, the results obtained with *Daphnia* and *Drosophila* point to similar molecular patterns of Dscam. The gene does not have high nucleotide diversity in both *Da. magna* and *Dr. melanogaster*. Instead, Dscam diversity is generated by alternative splicing of duplicated exons (more than 13000 and 30000 protein isoforms can potentially be expressed in *Da. magna* and *Dr. melanogaster*, respectively) and there is selection to preserve the diversity caused by duplication and divergence. In both taxa, epitope II coding regions diverged more than the rest of the gene, but in *Drosophila* we could not show that this high substitution rate was due to adaptive evolution. Epitope II coding regions harbor an excess of non-synonymous polymorphism in relation to the divergence levels observed. This could be maintained by balancing selection but also be influenced by segregating slightly deleterious mutations as discussed previously, which would suggest lower constraints on this part of the Dscam molecule.

Nevertheless, some of the segregating epitope II amino acids in both *Da. magna* and *Dr. melanogaster* populations might considerably change the binding capacities of the epitope (Fig. 2). In *Da. magna* arginine and glycine (exon 6.7) and in *Dr. melanogaster* arginine and methionine (exon 6.24) or asparagine and lysine (exon 6.39). In the case of the arginine polymorphism, the amino acid variants have exactly the same position in the epitope in both taxa in non-orthologous exons (Fig. 2). Furthermore, at this position glycine is a hallmark amino acid of many Ig domains [52] which corroborates the idea that this polymorphism might not be neutral. In *Da. magna* the arginine/glycine polymorphism showed an intermediate-frequency polymorphism with 54% of the analyzed individuals being homozygous for glycine, 30% being homozygous for arginine, and 17% being heterozygous across different populations. Both *Da. lumholtzi* and *Da. pulex* have glycine at this site.

Epitopes II are formed by the interception of two interstrand loops belonging to Ig2 and Ig3 domains (Fig. 2). This resembles "complementary determining regions" of T cell receptors or antibodies of the Immunoglobulin superfamily that, respectively, bind peptides or native antigenic determinants from pathogens (Fig. 2). A similar epitope in hemolin, a molecule involved in immunity in leptidopterans, has been suggested to harbor a similar region involved in bacterial lipopolysaccharide binding [53]. These and other structural similarities constitute circumstantial evidence for an involvement of Dscam in immunity, yet the molecular patterns we have found are not unequivocal.

Genes of the immune system involved in recognition, such as MHC, present hallmarks of long-term balancing selection; elevated levels of synonymous diversity and deeply diverged, trans-specific alleles. However, such strong patterns are not found in Dscam. It remains a challenge in the field of arthropod immunology to uncover the underlying mechanisms of the Dscam function. Expression by effector cells of the immune system such as hemocytes, is not in itself a guarantee of an involvement in immune recognition. Dscam diversity could play there a role similar to that played in neurons, controlling interactions between hemocytes inside the body.

REFERENCES

1. Watson LF, Püttmann-Holgado FT, Thomas F, Lamar DL, Hughes M, et al. (2005) Extensive diversity of Ig-superfamily proteins in the immune system of insects Science 309: 1874-1878

2. Chou PH, Chang HS, Chen IT, Lin HY, Chen YM, et al. (2009) The putative invertebrate adaptive immune protein Litopenaeus vannamei Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. Developmental and Comparative Immunology 33: 1258-1267.

3. Brites D, McTaggart S, Morris K, Anderson J, Thomas K, et al. (2008) The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Molecular Biology and Evolution 25: 1429-1439.

4. Dong Y, Taylor HE, Dimopoulos G (2006) AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biol 4: e229-.

5. Chen BE, Kondo M, Garnier A, Watson FL, Püettmann-Holgado R, et al. (2006) The Molecular Diversity of Dscam Is Functionally Required for Neuronal Wiring Specificity in *Drosophila*. Cell 125: 607-620.

6. Hattori D, Millard SS, Wojtowicz WM, Zipursky SL (2008) Dscam-Mediated Cell Recognition Regulates Neural Circuit Formation. Annual Review of Cell and Developmental Biology 24: 597-620.

7. Watthanasurorot A, Jiravanichpaisal P, Liu H, Söderhäll I, Söderhäll K (2011) Bacteria-iduced Dscam Isoforms of the crustacean, *Pacifastacus leniusculus*. PLoS Pathog 7: e1002062.

8. Vlisidou I, Dowling AJ, Evans IR, Waterfield N, ffrench-Constant RH, et al. (2009) *Drosophila* embryos as model systems for monitoring bacterial infection in real time. PLoS Pathog 5: e1000518.

9. Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J-h, Walz T, et al. (2007) Structural basis of Dscam

isoform specificity. Nature 449: 487-491.

10. Obbard DJ, Welch JJ, Kim KW, Jiggins FM (2009) Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet 5: e1000698.

11. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, et al. (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity Cell 101: 671-684.

12. Millard SS, Flanagan JJ, Pappu KS, Wu W, Zipursky SL (2007) Dscam2 mediates axonal tiling in the *Drosophila* visual system. Nature 447: 720-U714.

13. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

14. Miller M, Holder M, Vos R, Midford P, Liebowitz T, et al. (2009) The CIPRES Portals. CIPRES.

15. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16: 404-405.

16. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25: 4876-4882.

17. Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20: 426-427.

18. Stephens M, Smith NJ, Donnelly P (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data - Reply. American Journal of Human Genetics 69: 912-914.

19. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526-538.

20. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

21. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599.

22. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

23. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917.

24. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679.

25. Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Molecular Biology and Evolution 22: 2375-2385.

26. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11: 725-736.

27. Muse SV, Clark AG, Thomas GH (1997) Comparisons of the nucleotide substitution process among repetitive segments of the alpha- and beta-spectrin genes. Journal of Molecular Evolution 44: 492-500.

28. Yang ZH, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Molecular Biology and Evolution 19: 49-57.

29. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573.

30. Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3: 201-212.

31. Yang ZH, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites

along specific lineages. Molecular Biology and Evolution 19: 908-917.

32. Zhang JZ, Nielsen R, Yang ZH (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22: 2472-2479.

33. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39: 121-152.

34. Haag CR, McTaggart SJ, Didier A, Little TJ, Charlesworth D (2009) Nucleotide polymorphism and within-gene recombination in Daphnia magna and D. pulex, two cyclical parthenogens. Genetics 182: 313-323.

35. Little T, Colbourne JK, Crease T (2004) Molecular evolution of *Daphnia* immunity genes: polymorphism in a *gram-negative binding protein* gene and □□*Macroglobulin* gene. Journal of Molecular Evolution 59: 498-506.

36. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652-654.

37. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

38. Charlesworth J, Eyre-Walker A (2008) The McDonald-Kreitman test and slightly deleterious mutations. Molecular Biology and Evolution 25: 1007-1015.

39. Innan H (2009) Population genetic models of duplicated genes. Genetica 137: 19-37.

40. Lee C, Kim N, Roy M, Graveley BR (2009) Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. Rna 16: 91-105.

41. Takahata N, Nei M (1990) Allelic Genealogy Under Overdominant and Frequency-Dependent Selection and Polymorphism of Major Histocompatibility Complex Loci. Genetics 124: 967 -978.

42. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 400: 667-671.

43. Tennessen JA, Blouin MS (2008) Balancing Selection at a Frog Antimicrobial Peptide Locus: Fluctuating Immune Effector Alleles? Molecular Biology and Evolution 25: 2669 -2680.

44. Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. Plos Genetics 2: 379-384.

45. Rottschaefer SM, Riehle MM, Coulibaly B, Sacko M, Niare O, et al. (2011) Exceptional diversity, maintenance of polymorphism, and recent directional selection on the APL1 malaria resistance genes of Anopheles gambiae. PLoS Biol 9: e1000600.

46. Obbard DJ, Callister DM, Jiggins FM, Soares DC, Yan G, et al. (2008) The evolution of TEP1, an exceptionally polymorphic immunity gene in Anopheles gambiae. BMC Evol Biol 8: 274.

47. Kreitman M, Di Rienzo A (2004) Balancing claims for balancing selection. Trends in Genetics 20: 300-304.

48. Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA 86: 958-962.

49. Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America 98: 2509-2514.

50. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167-170.

51. Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution 18: 1585-1592.

52. Lefranc M-P, Lefranc G (2001) The Immunoglobulin Facts Book. London: Academic Press. 457 p.

53. Su XD, Gastinel LN, Vaughn DE, Faye I, Poon P, et al. (1998) Crystal structure of hemolin: A horseshoe shape with implications for homophilic adhesion. Science 281: 991-995.

SUPPLEMENTARY MATERIAL

TABLE S1 Non-synonymous polymorphisms and non-synonymous divergence in the duplicated exons of Dscam in *Daphnia.*

Exon[a]	Codon [b]	State[c]	AA[d]	Frequency (%)[e]
4.1	19	P	A/T	96.4
4.1	44 (II)	D	N/S	
4.2	90	P	E/D	96.4
4.2	100 (II)	D	N/T	
4.3	107	P	T/N	92.80
4.3	111	D	L/I	
4.3	135 (II)	D	I/T	
4.6	211 (II)	P	D/A	96.4
4.6	215 (II)	D	T/S	
4.6	218 (II)	D	P/Q	
4.7	243	P	A/V	96.4
4.7	264 (II)	D	G/S	
4.7	275 (II)	P	T/R	92.80
4.8	294	D	A/T	
4.8	317 (II)	D	G/D	
6.6	38	D	F/N	
6.6	39	D	F/N	
6.6	62 (II)	D	I/A	
6.6	63	P	S/F	93.75
6.6	78	D	F/Y	
6.7	84	P	A/S	93.75
6.7	102 (II)	P	G/R	71.8
6.7	103 (II)	P	M/I	93.75
6.1	75	P	F/Y	87.5
6.12	81	P	P/S	93.75
6.12	101[i] (II)	D	F/S/T	

[a] Array and exon numbering as in [3].

[b] Codon numbering within each exon. (II) indicates that the codon is in epitope II. [i] and [ii] refer respectively to nucleotides 658 and 659 in the same codon.

[c] P indicates a polymorphism within *Da. magna*, D a fixed difference between *Da. magna* and *Da. lumholtzi*, and P/D a polymorphic site within *Da. magna* at which *Da. lumholtzi* has a third amino acid.

[d] The first amino acid corresponds to the more common allele in the case of polymorphic (P and P/D sites). The last amino acid designates the one present in *Da. lumholtzi* (D and P/D sites).

[e] Frequency of the most common allele.

TABLE S2 Random sites model [23] likelihood ratio tests (LRT) for positive selection at MHC Class I locus B in six primate species. One allele per species was randomly chosen from Genebank (HQ231327.1 *Homo sapiens*, DQ026306.1 *Gorilla gorilla*, CR860073.1 *Pongo abelii*, AAB08074.1 *Hylobates lar*, AAY59437.1 *Pan troglodytes*, AAA50178.1 *Pan paniscus*). This analysis was done to assess the power of the random site model tests in our analysis of the *Drosophila* data, According to the results, the amino acid variation observed between the orthologous MHC alleles was more likely explained by neutral evolution (i.e., no significant signs of positive selection were found), which suggests that our site model analysis is not very powerful at detecting diversifying selection.

Model	LRT	Parameters [a]
M1a vs. M2a	$\chi 2=3.06$ df =2 $p=0.2$	M1a: ω_0=0 (71%) ω_1=1 (29%) M2a: ω_2=2 (21%)
M7 vs. M8	$\chi 2=3.1$ df=2 $p=0.2$	M7: p=0.005; q=0.011 M8: 'p =4.66, 'q=88 ω=2 (20%)

[a] ω_0, ω_1, ω_2 indicate the estimated values of ω under the conditions of each model; M1a: $0<\omega_0<1$, $\omega_1=1$; M2a adds to M1a $\omega_2>1$, which is estimated from the data; within brackets is the proportion of sites estimated to be in each category of ω. In M7, $0\leq\omega\leq1$ and p and q are parameters of the beta distribution. M8 adds one extra class of sites $\omega\geq1$ to M7.

TABLE S3 Non-synonymous polymorphisms in epitope II regions of array 6 exons in *Dr. melanogaster*. Shown are only polymorphisms at which the overall frequency of the rarer allele exceeds 0.15.The amino acids present at the orthologous codons in other *Drosophila* species is shown as well.

		Codon[a]						
Species	Population	65	9502	1027	1109	1547	1598	1625
Dr. melanogaster	Athens	S/G	R	P/L	A/S	N/K	I/S	A/V
	Florida	S/G	R	P/L	A/S	N/K	I/S	A/V
	French Polynesia	S/G	R	P/L	A/S	N	S	A/V
	Gabon	S/G	R/M	P	A	N/K	I/S	A
	Japan	S/G	R/M	P/L	A/S	N/K	I/S	A/V
	Kenya	S/G	R	P/L	A/S	N/K	I/S	A/V
Ancestral		G	R	P	A	N	S	A
Dr. simulans		G	R	A	A	K	S	A
Dr. sechellia		G	R	P	A	K	S	A
Dr. yacuba		G	R	A	A	K	S	n.o.

[a] Polymorphism data and codon numbering from [10]. n.o. indicates that no orthologous exon was found in this species.

Figure S1 Array 4 (A) and array 6 (B) partitions of epitope I and epitope II in *Da. magna*. Polymorphic positions are indicated by amino acids with the size of the letter being proportional to the frequencies of each amino acid. The colors represent the chemical properties of amino acids: polar (green), basic (blue), acidic (red) and hydrophobic (black). This figure was created with WebLogo (http://weblogo.berkeley.edu/logo.cgi).

A)

B)

Figure S2 Sliding window analysis across array 6 exons of the ratios of nonsynonymous nucleotide diversity π_a to synonymous nucleotide diversity π_s in *Da. magna* and of nonsynonymous divergence K_a to synonymous divergence K_s ratio between *D. magna* and *D. lumholtzi*. The sliding window analysis was done with DNAsp using a 50 bp window length with a 10 bp step size. The intron/exon boundaries as well as the locations of epitopes I (white bars, black dots) and epitopes II (grey bars) are indicated below the x-axis..

.

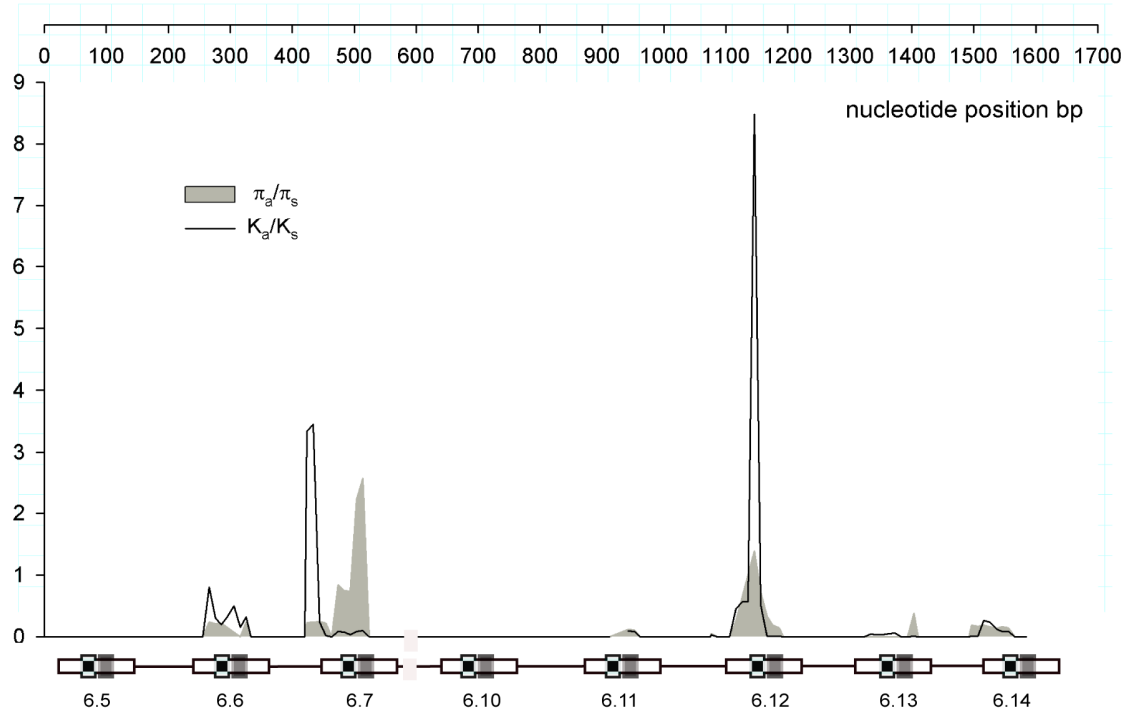
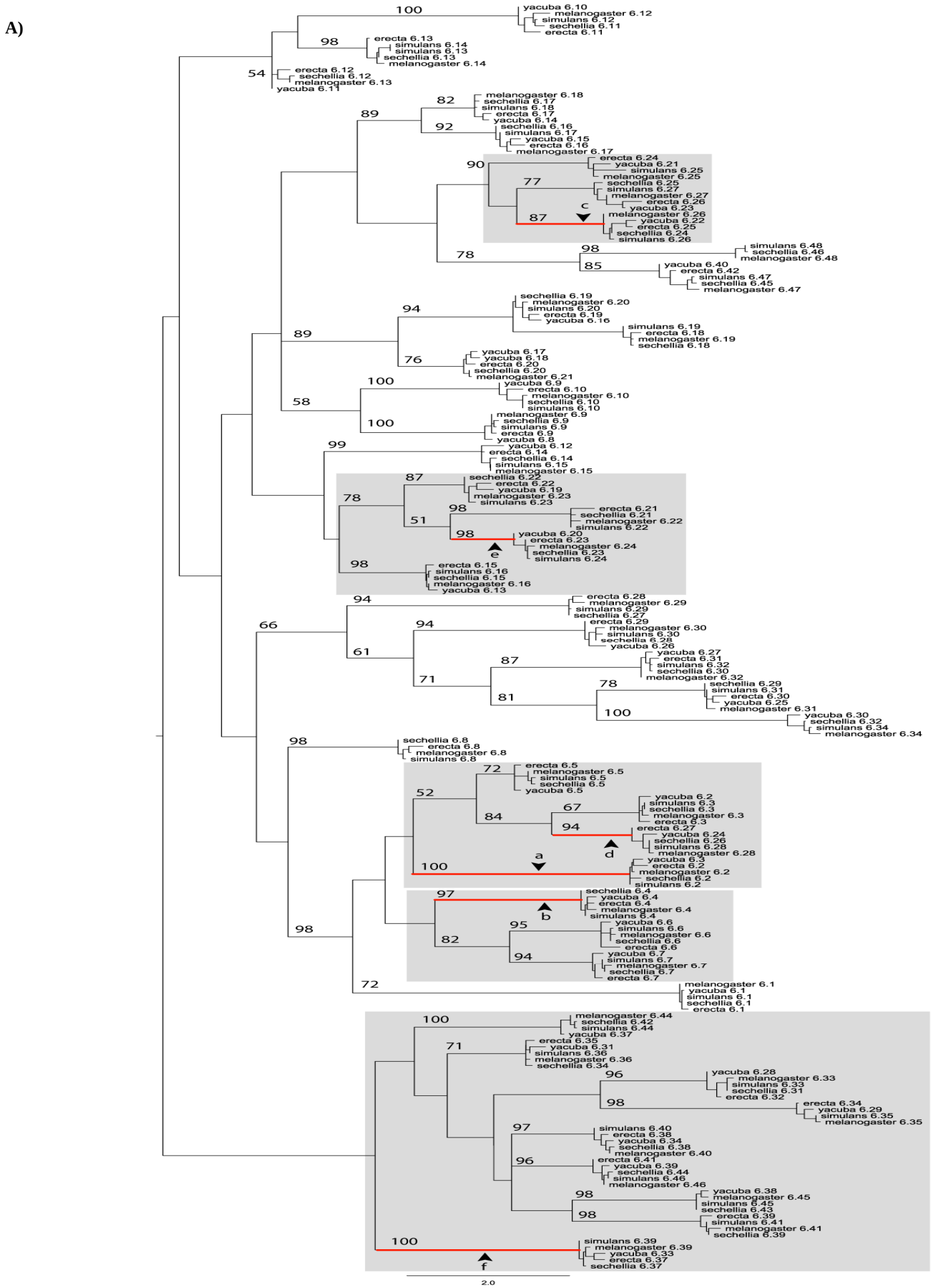


Figure S3 A) Maximum likelihood tree of array 6 exons in the melanogaster subgroup including orthologous and paralogous exons**.** Support values at nodes are bootstrap values (100 bootstrap replicates). Branch length estimates the expected number of nucleotide substitutions per codon using the one-ratio model, and the tree topology and branch lengths were used to fit different models. The tree is rooted for convenience at the midpoint but all analyses were done with an unrooted topology. Red branches with arrows indicate branches for which the presence of aminoacid sites that evolved with $\omega>1$ was tested using branch-site models implemented in PAML [31,32]. The branches chosen were the ones leading to duplicated exons where we detected an excess of non-synonymous polymorphism in *Dr. melanogaster* using McDonald-Kreitman tests. the PAML tests used smaller subtrees (grey boxes). B) Schematic representation of branch models. We used these models to test whether selection changed after duplication, that is whether orthologous and paralogous branches differ in ω (model R2). The null model R1 assumes that all branches in the tree have the same ω.

A)


yacuba 6.10
melanogaster 6.12
simulans 6.12
sechellia 6.11
erecta 6.11
erecta 6.13
simulans 6.14
simulans 6.13
sechellia 6.13
melanogaster 6.14
erecta 6.12
sechellia 6.12
melanogaster 6.13
yacuba 6.11
melanogaster 6.18
sechellia 6.17
simulans 6.18
erecta 6.17
yacuba 6.14
sechellia 6.16
simulans 6.17
yacuba 6.15
erecta 6.16
melanogaster 6.17
erecta 6.24
yacuba 6.21
simulans 6.25
melanogaster 6.25
sechellia 6.25
simulans 6.27
melanogaster 6.27
erecta 6.26
yacuba 6.23
melanogaster 6.26
yacuba 6.22
erecta 6.25
sechellia 6.24
simulans 6.26
c
simulans 6.48
sechellia 6.46
melanogaster 6.48
yacuba 6.40
erecta 6.42
simulans 6.47
sechellia 6.45
melanogaster 6.47
sechellia 6.19
melanogaster 6.20
simulans 6.20
erecta 6.19
yacuba 6.16
simulans 6.19
erecta 6.18
melanogaster 6.19
sechellia 6.18
yacuba 6.17
yacuba 6.18
erecta 6.20
sechellia 6.20
melanogaster 6.21
yacuba 6.9
erecta 6.10
melanogaster 6.10
sechellia 6.10
simulans 6.10
melanogaster 6.9
sechellia 6.9
simulans 6.9
erecta 6.9
yacuba 6.8
yacuba 6.12
erecta 6.14
sechellia 6.14
simulans 6.15
melanogaster 6.15
sechellia 6.22
erecta 6.22
yacuba 6.19
melanogaster 6.23
simulans 6.23
erecta 6.21
sechellia 6.21
melanogaster 6.22
simulans 6.22
yacuba 6.20
erecta 6.23
melanogaster 6.24
sechellia 6.23
simulans 6.24
e
erecta 6.15
simulans 6.16
sechellia 6.15
melanogaster 6.16
yacuba 6.13
erecta 6.28
melanogaster 6.29
simulans 6.29
sechellia 6.27
erecta 6.29
melanogaster 6.30
simulans 6.30
sechellia 6.28
yacuba 6.26
yacuba 6.27
erecta 6.31
simulans 6.32
sechellia 6.30
melanogaster 6.32
sechellia 6.29
simulans 6.31
erecta 6.30
yacuba 6.25
melanogaster 6.31
yacuba 6.30
sechellia 6.32
simulans 6.34
melanogaster 6.34
sechellia 6.8
erecta 6.8
melanogaster 6.8
simulans 6.8
erecta 6.5
melanogaster 6.5
simulans 6.5
sechellia 6.5
yacuba 6.5
yacuba 6.2
simulans 6.3
sechellia 6.3
melanogaster 6.3
erecta 6.3
erecta 6.27
yacuba 6.24
sechellia 6.26
simulans 6.28
melanogaster 6.28
d
a
yacuba 6.3
erecta 6.2
melanogaster 6.2
sechellia 6.2
simulans 6.2
sechellia 6.4
yacuba 6.4
erecta 6.4
melanogaster 6.4
simulans 6.4
b
yacuba 6.6
simulans 6.6
melanogaster 6.6
sechellia 6.6
erecta 6.6
yacuba 6.7
simulans 6.7
melanogaster 6.7
sechellia 6.7
erecta 6.7
melanogaster 6.1
yacuba 6.1
simulans 6.1
sechellia 6.1
erecta 6.1
melanogaster 6.44
sechellia 6.42
simulans 6.44
yacuba 6.37
erecta 6.35
yacuba 6.31
simulans 6.36
melanogaster 6.36
sechellia 6.34
yacuba 6.28
melanogaster 6.33
simulans 6.33
sechellia 6.31
erecta 6.32
erecta 6.34
yacuba 6.29
simulans 6.35
melanogaster 6.35
simulans 6.40
erecta 6.38
yacuba 6.34
sechellia 6.38
melanogaster 6.40
erecta 6.41
yacuba 6.39
sechellia 6.44
simulans 6.46
melanogaster 6.46
yacuba 6.38
melanogaster 6.45
simulans 6.45
sechellia 6.43
erecta 6.39
simulans 6.41
melanogaster 6.41
sechellia 6.39
simulans 6.39
melanogaster 6.39
yacuba 6.33
erecta 6.37
sechellia 6.37
f
2.0

B)

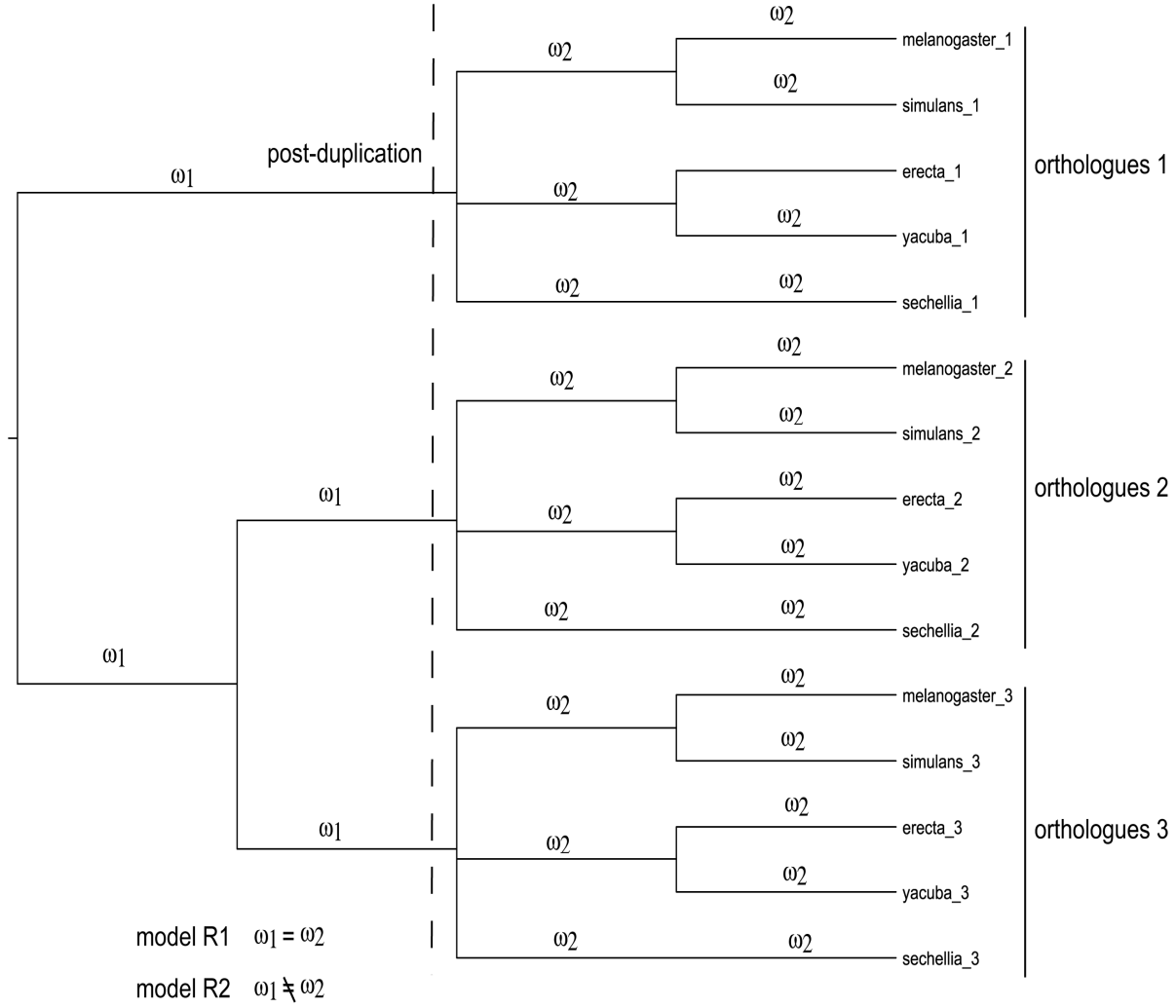

ω2
melanogaster_1
ω2
simulans_1
post-duplication
ω1
ω2
erecta_1
orthologues 1
ω2
yacuba_1
ω2
ω2
sechellia_1
ω2
melanogaster_2
ω2
ω2
simulans_2
ω1
ω2
erecta_2
ω2
ω2
yacuba_2
ω2
ω2
sechellia_2
ω1
ω2
melanogaster_3
ω2
ω2
simulans_3
ω2
ω1
erecta_3
ω2
ω2
yacuba_3
model R1 ω1 = ω2
ω2
ω2
sechellia_3
model R2 ω1 ≠ ω2

TABLE S4 Branch models and branch-site models applied to the exons of array in the melanogaster subgroup. Likelihood ratio test (LRT), parameter estimates (ω), and positively selected sites are shown. In branch-site models the branch of interest is called foreground branch (Fig. S3, red branches with arrows) and all the other branches in the tree are called background branches.

Models	LRT	*Parameters*	Positively selected sites[b]
Branch models			
One-ratio (R1) vs. Two-ratios (R2)	$\chi2$=46 df =1 p<0.001	ω_1=0.26 ω_2=0.094	
Branch-site models		*Parameters*[a]	
Foreground branch (**a**) vs. Background	$\chi2$=1.46 df =1 p=0.2	ω_0=0.07 ω_1=1 ω_{2a}B=0.07 ω_{2a}F=5.43 ω_{2b}F=5.43	10T**; 15 S*; 16 R*; 25 S**
Foreground branch (**b**) vs. Background	$\chi2$=0.38 df =1 p=0.55	ω_0=0.08 ω_1=1 ω_{2a}B =0.08 ω_{2a}F=2.32 ω_{2b}F=2.32	18 T*; 21 P**; 37 V**
Foreground branch (**c**) vs. Background	$\chi2$=0.09 df =1 p=0.8	ω_0=0.08ω_1=1 ω_{2a}B =0.08 ω_{2a}F =1 ω_{2b}F=1	
Foreground branch (**d**) vs. Background	$\chi2$=0 df =1 p=1	ω_0=0.02 ω_1=1 ω_{2a}B =0.02 ω_{2a}F =1 ω_{2b}F=1	
Foreground branch (**e**) vs. Background	$\chi2$=0 df =1 p=1	ω_0=0.05 ω_1=1 ω_{2a}B =0.05 ω_{2a}F =1 ω_{2b}F=1	
Foreground branch (**f**) vs. Background	$\chi2$=0 df =1 p=1	ω_0=0.08 ω_1=1 ω_{2a}B =0.08 ω_{2a}F =1 ω_{2b}F=1	

[a] Parameter estimates under the alternative models: ω_0:dN/dS<1; ω_1: dN/dS=1, ω_{2a}F= dN/dS >1 (alternative hypothesis) or dN/dS=1 (null hypothesis) on the foreground branch and dN/dS<1 on background branches,ω_{2a}B; ω_{2b}F=dN/dS >1 (alternative hypothesis) or dN/dS=1 (null hypothesis) on the foreground branch and dN/dS=1 on background branches [b] Sites inferred to be under positive selection at the 95% (*) or 99% (**) by Bayes Empirical Bayes analysis.

Table S5 Estimates of divergence between *Da. magna* and *Da. similis*, as well as McDonald Kreitman tests for the comparison between the two species. No polymorphisms were excluded for this analysis.

Gene region	*Da. magna vs Da. similis*							
	Divergence (k)			Fixed		Polymorphic		p[a]
	Ks	Ka	Ka/Ks	Syn	Nonsyn	Syn	Nonsyn	
Array 4 Total	0.094	0.011	0.117	21	8	4	6	0.12
Epitopes II	0.07	0.027	0.35	5	6	2	2	1

[a] p values are according to a two-tailed Fisher's exact test.

CHAPTER 4

DUPLICATION AND LIMITED ALTERNATIVE SPLICING OF DSCAM GENES FROM BASAL ARTHROPODS

Daniela Brites, Carlo Brena, Dieter Ebert and Louis Du Pasquier

manuscript

ABSTRACT The Dscam homologue of pancrustaceans is the most remarkable example known of how exon duplication and alternative splicing contribute to generate protein diversity. Here we describe for the first time Dscam homologues in the centipede *Strigamia maritima* and in the tick *Ixodes scapularis*, taxa that belong to two arthropod basal groups, the myriapods and chelicerates respectively. In both, Dscam diversified extensively by duplications of the whole Dscam gene and in some cases by duplications of exons coding for Immunoglobulin domain 7 (Ig7) and Ig8 but not of exons coding for half of Ig2 and Ig3 like in pancrustaceans. This resulted in the creation of a Dscam multigene family with many members in both *S. maritima* and *I. scapularis* which, according to our phylogenetic analysis share a common origin but expanded independently. We demonstrate furthermore that the mechanism of mutually exclusive AS known in pancrustaceans was already present *S. maritima* contributing to generate Ig7 diversity in both nervous and immune cells. That indicates that Dscam mutually exclusive AS and expression by hemocytes is not a derived character of pancrustaceans. Additionally, diversity caused by alternative splicing of the cytoplasmic domains of the receptor was also uncovered. We found evidence in both *S. maritima* and *I. scapularis* of extensive rearrangements among different Dscam paralogues and we propose that the highly variable Dscam gene of pancrustaceans evolved by recombination between Dscam paralogues with Ig7 coding exon duplications, from a common ancestor with more Dscam genes than any of the extant species of pancrustaceans. The convergent evolution of mechanisms to generate Dscam diversity in different arthropod groups suggests that the concomitant functional diversity created was important in the evolution of this very successful group.

INTRODUCTION

The Down syndrome cell adhesion molecule (Dscam) gene family is composed of several members related to other cell adhesion molecules (CAMs) like axonin, roundabout, etc, which are involved in the nervous system development (Shapiro, Love, and Colman 2007). The composition of the different Dscam members is relatively conserved among metazoa, consisting of 9(Ig)-4(FN)-Ig-2(FN) followed by a transmembrane domain and a less conserved cytoplasmic tail. Vertebrates and insects have paralogous Dscam members that resulted from whole gene duplications like DSCAM and DSCAM like (DSCAM-L) in vertebrates, and Dscam-L2, Dscam-L3 and Dscam-L4 in insects (Yamakawa et al. 1998; Schmucker et al. 2000; Agarwala et al. 2001; Millard et al. 2007). In the latter group, another homologue called Dscam, is the most remarkable example known of protein diversification by duplication and alternative splicing (AS) (Schmucker et al. 2000). In this member of the Dscam family certain exons duplicated extensively forming three arrays, that encode half of Ig2 and Ig3 domains, the complete Ig7 and two transmembrane domains (Schmucker et al. 2000) (Watson et al. 2005) (Fig. 1).

An exquisite form of mutually exclusive alternative splicing of the exon duplications ensures that only one exon per array is included in the mature mRNA (Schmucker et al. 2000; Graveley 2005; Kreahling and Graveley 2005; Olson et al. 2007). In this way, the *Drosophila melanogaster Dscam* gene has the potential to generate 19 008 different extracellular Dscam isoforms combined with two alternative transmembrane domains. Additionally, by alternative splicing four different cytoplasmic tails are used and hence, in total 152 064 different isoforms can be encoded in a single fly (Yu et al. 2009).

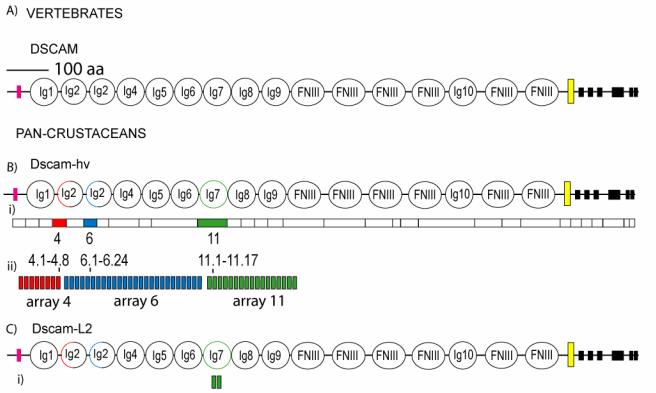


Figure 1 – Dscam domain representation; Ig- immunoglobulin domains; FNIII- fibronectin III domains. The yellow and black boxes represent the transmembrane and cytoplasmic domains. A) DSCAM of vertebrates represented by the homologue in *Homo sapiens* B) Dscam-hv of pancrustaceans represented by the homologue in *Daphnia magna*; ii) mRNA, each box corresponds to a constitutive exons and the colored boxes 4, 6 and 11, correspond to exons that are the result of mutual exclusive alternative splicing of arrays of duplicated exons which are present in three arrays, as indicated in ii); C) Dscam-L2 of pancrustaceans; i) two exons that are mutually exclusive alternatively spliced code for Ig7.

A homologue of this gene is also present in crustaceans with a similar organization but with only one transmembrane domain coding exon (Brites et al. 2008; Chou et al. 2009). For the sake of clarity we will designate hereafter this Dscam member of insects and crustaceans (pancrustaceans) as Dscam hypervariable (Dscam-hv). The mechanism of Dscam somatic

diversification described has not been observed in deuterostomes so far, except for the generation of two transmembrane forms in humans, but through a much simpler mechanism (Yamakawa et al. 1998).

Despite the differences, DSCAM and Dscam-hv are both involved in similar developmental processes controlling neural wiring (for a review see Hattori et al. 2008). Additionally, the diversity of Dscam-hv isoforms in pancrustaceas seems to play a role in the immune system (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006; Watthanasurorot et al. 2011). The silencing of the gene reduces the phagocytosis activity of hemocytes, infection by different pathogens induces different alternative splicing patterns of the molecule and different isoforms have different binding specificities to different bacteria (Watson et al. 2005; Watthanasurorot et al. 2011). Furthermore, Dscam-hv soluble forms circulate in the hemolymph of both insects and crustaceans suggesting that they could function as opsonins but with a function not yet fully elucidated (Watson et al. 2005; Watthanasurorot et al. 2011).

It has been generally assumed that the diversification of Dscam-hv has occurred in all arthropods (Crayton et al. 2006; Kurtz and Armitage 2006; Lee et al. 2009). Arthropods appeared approximately 600 million years ago and represent far more species than any other animal phyla (Budd and Telford 2009). The high diversity of living arthropod species is grouped in four taxa; insects, crustaceans, chelicerates and myriapods. Dscam in the latter two taxa has not been studied so far. Here we report on Dscam related genes in the tick *Ixodes scapularis*, a chelicerate, and in the centipede *Strigamia maritima*, a myriapode. We also studied the expression of one Dscam homologue in *Strigamia maritima*. This broadened the phylogenetic sampling of Dscam genes in arthropods and revealed interesting differences, but also similarities, among Dscam in the different arthropod groups which are relevant for understanding the evolutionary history this gene family.

MATERIAL AND METHODS

Gene recovery

The program tblastn was used to probe several genomes (Table S1) to search for Dscam related genes. We did first a general search using the whole Dscam-hv of *Drosophila melanogaster* and selected the most related genes based on amino acid similarity and domain architecture. Several architectural criteria were used non-exclusively; the Ig1 motif GxxxxC (where x stands for any amino acid and C refers to the first cysteine in the Ig domain) which is a distinctive signature of Dscam (in regular Ig domains G is at position -8 in relation to the cysteine referred); the presence of Ig1 to Ig4, which are domains that form a horse-shoe structure typical of Dscam and other related CAMs (Meijers et al. 2007); and the presence of

Ig10 in an intermediate position between the FNIII domains. Finally we looked for the transmembrane domains and cytoplasmic tails sequence similarities. In all Dscam related genes found we did a further search for duplicated exons using the Dscam-hv variable regions of Ig2, Ig3 and Ig7. All homologues were annotated by hand using the identity information and a prediction of the protein structure obtained with SMART (http://smart.embl-heidelberg.de) (Schultz et al. 1998; Letunic, Doerks, and Bork 2009).

Identification and annotation of the Dscam of Myriapodes and Chelicerata

The procedure described above was used to search for Dscam related genes in the genomes of *Ixodes scapularis* (http://www.vectorbase.org/index.php) and that of *Strigamia maritima* 24X scaffolding (http://www.hgsc.bcm.tmc.edu/collaborations/insects/strigamia/). In both taxa, several Dscam related genes were incomplete and/or did not correspond exactly to the Dscam canonical architecture 9(Ig)-4(FN)-(Ig)-2(FN) (Shapiro, Love, and Colman 2007). In our analysis we included only the members which we believed as not being the result of assembly mistakes. Each gene was named after the name of species to which it belongs followed by a number (Fig. S2 and Fig. S5). In this way, all *I. scapularis* and *S. maritima* Dscam homologues start with Is and Sm, respectively. We have furthermore scrutinized the EST data base available for *I. scapularis* to look for Dscam expression using the same blast procedure described above (http://iscapularis.vectorbase.org/SequenceData/EST/).

Phylogenetic reconstruction

Multiple alignments of amino acid sequences were built using CLUSTALW and edited through Jalview (Waterhouse et al. 2009). The G, W and C amino acids at certain positions are distinct features of Ig domains (Lefranc and Lefranc 2001) and were used as reference amino acids to correct the alignments manually. Phylogenetically conflicting regions of the alignments were eliminated following Gblocks selected blocks (Castresana 2000; Talavera and Castresana 2007) .The program ProTest 1.4 was used to estimate the amino acid substitution model and related the parameters that better describe the evolution of the aligned sequences (Drummond and Strimmer 2001; Guindon and Gascuel 2003; Abascal, Zardoya, and Posada 2005). This information was used to build protein phylogenies with both Bayesian and Maximum Likelihood (ML) methods, using MrBayes 3.1.2 and RAxML (Stamatakis 2006), respectively. For the Bayesian analysis we used a gamma rate distribution estimated from our dataset and a burn-in equal to 1/10 the number of generations; after the burn-in phase every 100th tree was saved. Two parallel Markov chains were run simultaneously in each of two runs. Tree length, log-likelihood score and alpha value of the gamma distribution were examined prior

to the termination of MrBayes to ensure that all parameters had reached stationarity. To access whether the MCMC of the two runs converged we used AWTY (Nylander et al. 2008) for plotting the posterior probabilities of all splits for the two runs and increased the number of generations when necessary. For the ML analysis we run RAxML through the Cipres Portal (Miller et al. 2009) with at least 1000 bootstrap replicates.

To determine the homology of Dscam related genes found in basal metazoan groups, we estimated phylogenies of 42 proteins including Dscam and other proteins from the CAM family whose Ig1 to Ig4 domains form a horse-shoe structure (Table S1). This phylogeny was rooted using the sequence of human NCAM (Neural cell adhesion molecule), a immunoglulin superfamily CAM that does not form a horse-shoe tertiary structure.

The relationship between all Dscam homologues representative of major metazoan clades was reconstructed by estimating phylogenies based on aligned Dscam sequences of Ig2 to FNIII-2 domains given that Ig1 was not found in many cases. In order to include incomplete Dscam homologues of *Ixodes* with multiple exons coding for Ig7 and Ig8, we estimated phylogenies based on Ig8 to FNIII-2 domains. To trace the origins of Ig7, phylogenetic trees of all Ig7 domains of Dscam and Dscam-L of all arthropods and deuterostomes were produced. Due to the high number of exons analysed (177) we present only the results of the confident monophyletic groups of exons found (exons that shared their most recent common ancestor with 0.95 posterior probability and that were grouped in more than 60% bootstrap replicates in the Bayesian and ML analysis, respectively).

Strigamia maritima dissections, RNA extraction and cDNA synthesis

Adult individuals of *Strigamia maritima* were sampled near Bora, Scottland and kept alive at 4°. RNA was extracted from whole-body, hemocytes and heads using Trizol (INVITROGEN) following manufacturer instructions. In the case of hemocytes and heads, to increase RNA yield, RNA samples were precipated overnight in isopropanol at -80° with 5 µg of RNAse free glycogen added (INVITROGEN). Hemocytes were obtained by cutting the individuals in several sections and withdrawing the hemolymph by capillary action using microcapillary glass tubes (Harvard apparatus GC100TF-10). To check the expression of Dscam in the nervous system, the heads from the same individuals were used for RNA extraction. All material was immediately stored in RNA*later* (Ambion) solution.

To obtain the 5' leader region of the *Sm35* gene of *S. maritima*, we used SMART technology (SMART^{tm} RACE cDNA Amplification Kit, CLONTECH) on mRNA samples extracted from whole-body following the instruction of the manufacturer and specific reverse primer annealing to Ig3.

The expression of the duplicated exons of *Sm35* coding for Ig7 was investigated by sequencing RT-PCR amplicons obtained with primers specific to Ig6 and Ig8 coding exons. For this purpose the One Step PCR kit (QUiagen) was used to perform a multiplex PCR with the *Sm35* specific primers and primers specific to actin to serve as positive controls. All PCR products were cloned in the pCR 2.1- TOPO vector (Invitrogen) and sequenced with traditional Sanger sequencing.

RESULTS

The Dscam family within the Immunoglobulin superfamily CAMs

We found Dscam related genes in metazoan basal groups such demosponges (*Amphimedon queenslandica*), cnidarians (*Nematostella vectensis*) and a placozoan (*Tricoplax adhaerens*) (Table S1). These genes do not encode proteins with canonical Dscam architectures. To investigate whether they belong to the Dscam family we built a phylogeny including those, other metazoan Dscam proteins and some other cell adhesion molecules (CAMs) from the immunoglobulin superfamily whose first four Ig domains, like in Dscam, form a horse-shoe structure (roundabout, axonin, L1CAM and hemolin). Most Dscam genes formed relatively well supported clades and most likely have a monophyletic origin although the latter could not be recovered with statistical support (Fig. 2). The same is true for roundabout and axonin, molecules which are used by the nervous system and to which the gene of *T. adhaerens* is most closely related. We could not recover with confidence the relationship between the genes from *A. queenslandica* and three of the genes in *N. vectensis* and the remaining CAMs (Fig. 2). All blasted significantly to Dscam but did not form any well supported clade in our analysis (Fig. 2). The position of *N. vectensis* gene Nv_1 is unclear based on the phylogenetic relationships estimated using the first four Ig domains of the molecule. Yet, if the phylogeny is based on region comprising Ig8 to FNIII-2 domains, Nv_1 forms a well supported clade with the human Dscams (Fig. S1) reflecting the similarity of Dscam with vertebrate Dscam (approximately 30% similarity, E values between e^{-171} and e^{-179}). Furthermore, their cytoplasmic tails also share similar SH2, ITIM and polyproline motifs (data not shown) indicating that they use similar signaling pathways. In subsequent analysis of the Dscam gene family, the gene *Nv_1* was used as an outgroup sequence.

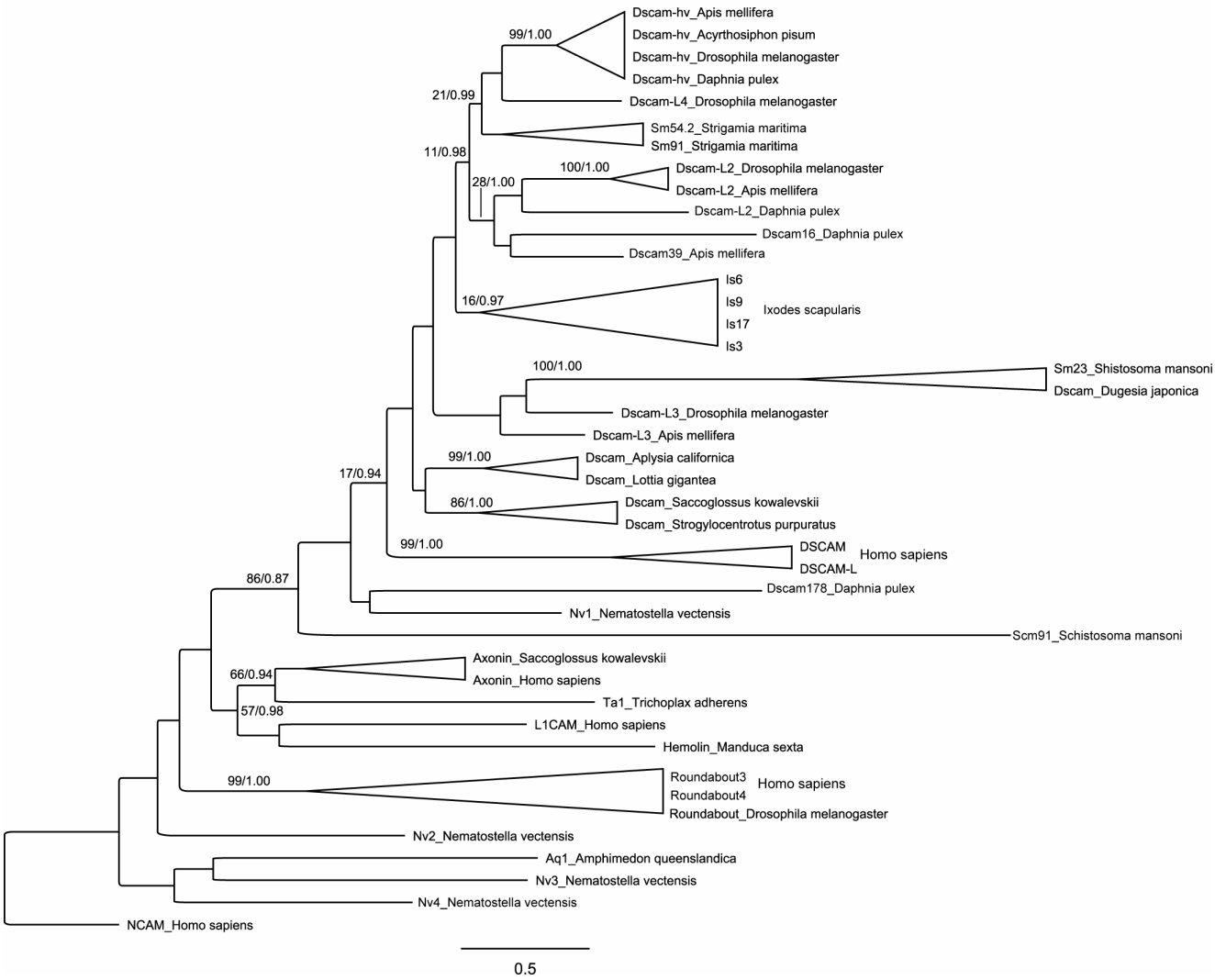


Figure 2- Maximum likelihood topology 42 CAMs whose first four Ig domains form a horse-shoe tertiary structure. Support values at nodes are bootstrap values relative to 1000 replicates (left value) and posterior probabilities (right value) when higher than 60% and/or than 0.95, respectively. The tree is rooted with the human NCAM, a CAM which does not form a horse-shoe structure

Diversification of Dscam in chelicerates and myriapodes

Extracellular domain diversification by gene and domain duplication

A very high number of Dscam related genes was found in both *I. scapularis* and *S. maritima* genomes. None exhibits internal duplications of exons coding for Ig2 and Ig3 domains like the Dscam-hv gene of pancrustaceans but a few genes have duplications of exons coding for Ig7. The purpose of the present study was not an exhaustive description of all the Dscam genes in *S. maritima* and *I. scapularis*, but an analysis of relevant comparative aspects with Dscam genes from other taxa. For that reason we have

annotated only a fraction of the Dscam genes present in the genome of those organisms. Although all statements about absence of genes or domains have to be taken carefully, especially in the case of *I. scapularis* for which many of the analyzed genomic scaffolds were interrupted by undetermined sequences, we are fairly confident in our claim that in the current genome assemblies there are no arrays of duplicated exons coding for Ig2 and Ig3 like in the canonical Dscam-hv of pancrustaceans.

Strigamia maritima In the myriapod *S. maritima* we found a high number of Dscam related genes present in the current genome assembly (approximately 50 hits with $E>10^{-4}$, depending on which Dscam domains were used as query sequence). The majority of genes are strongly similar to Dscam, although some are incomplete or do not correspond to the canonical structure. An equivalent of the arrays of exons coding for half of Ig2 and Ig3 domains present in pancrustaceans was not found. In contrast several genes present arrays of duplicated exons coding for Ig7; genes *Sm35, Sm54.1 and Sm62.2* have four duplicated exons, genes *Sm62.1* and *Sm55* have three and genes *Sm91* and *Sm546* have two Ig7 coding exon duplications (Fig. S2A). The phylogenetic relationship between the exon duplicates indicates that they were probably already present before the genes duplicated as they are more similar between genes than within each gene (Fig. S3). Assuming that this is true, one would expect that those Ig7 domains have similar amino acid divergence compared to the remaining ectodomains of those paralogous Dscam genes. Interestingly, the aminoacid sequences of the duplicated Ig7 domains are less divergent than the remaining ectodomains (Fig. S4), suggesting that they might be under gene conversion or recombination.

Ixodes scapularis We found 27 genes with strong similarity to Dscam although none exhibits the exact configuration of a canonical Dscam, generally lacking the third and fourth FNIII domains and the tenth Ig domain (Fig. S5). Fifteen almost complete homologues could be reconstructed (Fig. S5) and analyzed but the number of contigs with Dscam related genes amounts in total to 56, often containing strongly related but single Dscam domains. In the current assembly we did not find exon duplicated arrays coding for half of Ig2 and Ig3 like in the Dscam-hv of pancrustaceans. Instead we found four genes *Is27*, *Is28*, *Is29* and *Is53*, each with several duplications of exons coding for Ig7 and Ig8 (Fig. 3A). The multiple exons coding for Ig7 and Ig8 are in alternate positions in the genome, a feature not observed in any other Dscam gene (Fig. 3A). The exon and intron structure of these genes sugests that they could be alternatively spliced but no related ESTs were found.

The genes *Is27*, *Is28* and *Is29* are located in the same contig separated approximately by 1900 bp. Genes *Is28* and *Is29* are duplicates of each other, whereas the origin

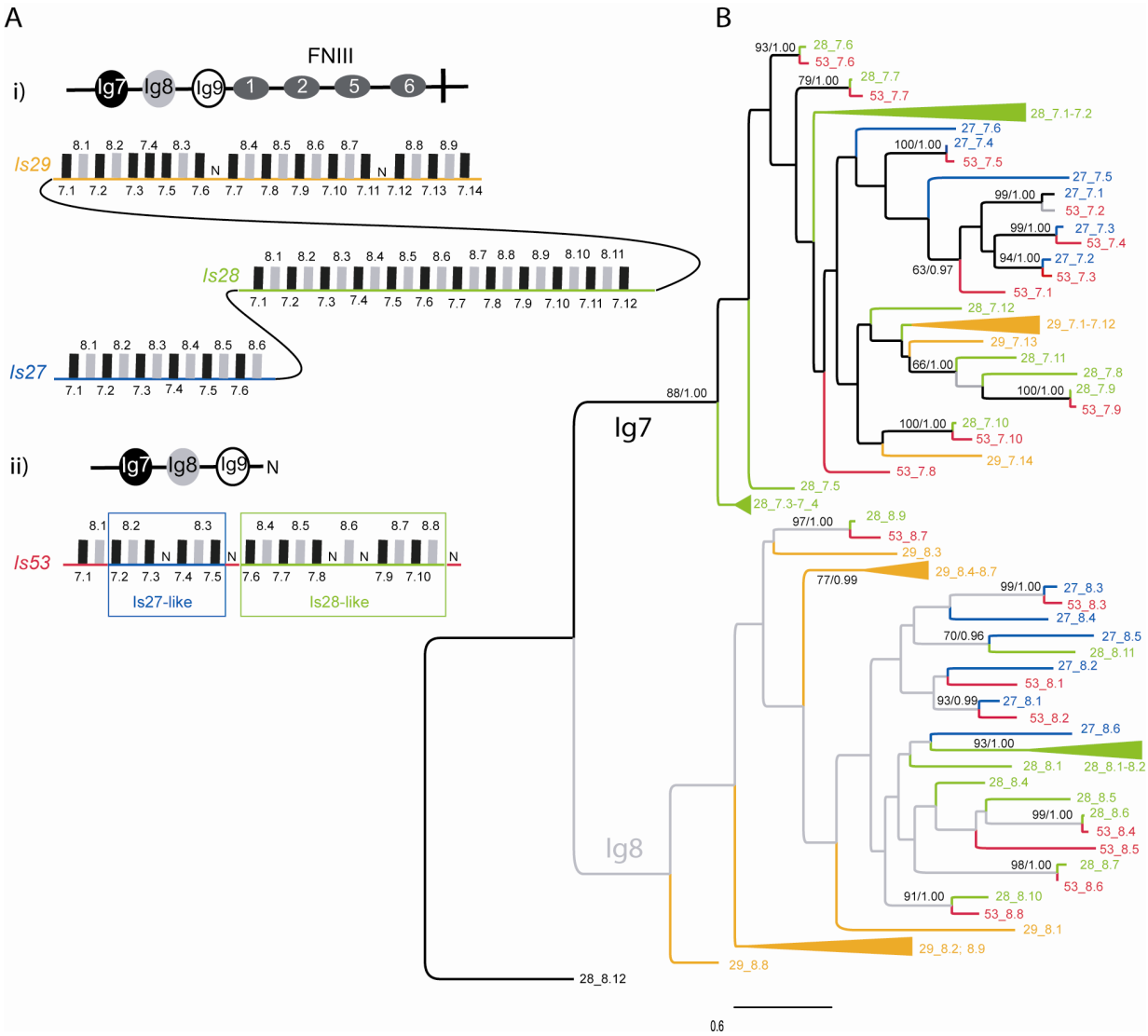


Figure 3 A *Ixodes scapularis* Dscam homologues with duplicated exons coding for Ig7 and Ig8 i) protein reconstruction coded by genes *Is27*, *Is28* and *Is29* which are all adjacent in the same contig. ii) protein reconstruction coded by *Is53*. Bellow each reconstruction is the representation of the alternative exons of each gene coding for Ig7 (black boxes) and ig8 (grey boxes). N represents undetermined sequence. **B** Maximum likelihood topology of the duplicated exons coding for Ig7 (black branches) and Ig8 (grey branches) in *I. scapularis* Dscam homologs *Is27* (blue branches), *Is28* (green branches), *Is29* (orange branches) and *Is53* (red branches). Support values at nodes are bootstrap values relative to 1000 replicates (left value) and posterior probabilities (right value) when higher than 60% and/or than 0.95, respectively. The tree is rooted for convenience with exon 8.12 from gene *Is28* because this exon has the lowest aminoacid similarity relative all other exons in the tree. Monophyletic clades of exons were collapsed for convenience.

of Is27 is not possible to elucidate (Fig. S1). Nevertheless, an contrarily to the Ig7 duplications in *S. maritima*, the multiple duplications coding for Ig7 and Ig8 seem to have occurred independently in the three genes, since paralogous exons within each gene are more similar to each other than to paralogous exons in the other genes (or they diverged so extensively

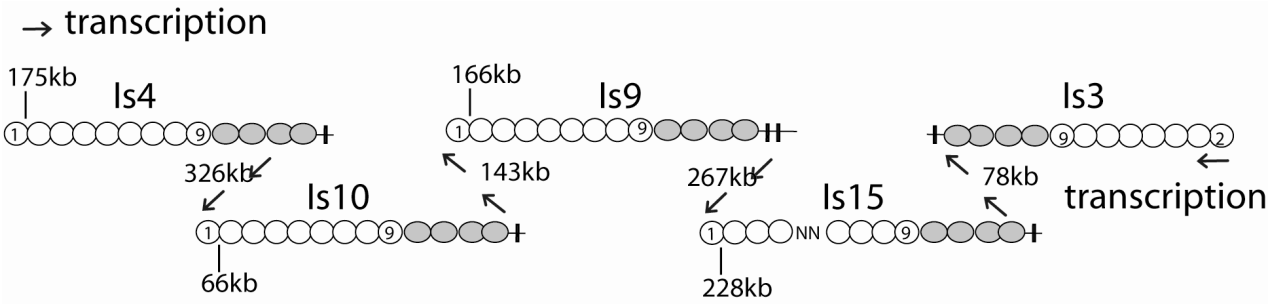


Figure 4 *I. scapularis* reconstructions of Dscam homologues present in contig 92235. Ig domains are represented by open circles and FN domains by grey ellipses. The genomic regions between these genes are represented by arrows and its size is indicated. The size of the genomic regions between the exons that code for Ig1 and Ig2 are indicated as well. NN indicates that the sequence was undetermined

that a common origin cannot be discerned) (Fig. 3B). The only exceptions to this are exons coding for *Is27* Ig8.5 and *Is28* Ig8.11 (Fig. 3B). Contrastingly, the gene *Is53* has a chimerical arrangement originated from a whole duplication of the *Is27* region containing exons 7.1 to 7.5 and a whole duplication of the *Is28* region containing exons 7.6 to 7.10 (Fig. 3A & B). The conservation of amino acids is very strong between *Is53* and *Is27* and *Is28* but not at the nucleotide sequence, excluding the possibility that this is an artifact of the assembly. Additionally there are no pseudoexons suggesting that these are functional genes. Genes *Is15*, *Is4*, *Is9*, *Is10* and *Is3* were also found to be physically close in the genome and all are transcribed in the same direction, except *Is3* (Fig. 4). The phylogenetic relationships among these genes are mostly unresolved except for *Is3* which is most closely related to *Is26*, a gene present in a different genomic region (Fig. S1, Fig. 7).

Dscam diversification by alternative splicing in myriapod

In order to investigate whether the mechanism of mutually exclusive alternative splicing was already present in a Dscam member of *S. maritima* with internal duplicated exons coding for Ig7, we cloned and sequenced RT-PCR amplified fragments of the gene *Sm35* containing the duplicated exons obtained from RNA from whole single animals. We found transcripts containing Ig7 duplicated exons expressed in many possible ways; the four duplicated Ig7 coding exons can expressed in a mutually exclusive alternatively spliced fashion just like in Dscam-hv. Moreover, two alternative exons can be retained or Ig7 coding exons can be skipped all together (Fig. 5). This suggests that the mechanism of mutually exclusive alternative splicing of the Dscam-hv gene has evolved initially in the array of exon duplications coding

for Ig7 and it was already present in the ancestor of the pancrustaceans.

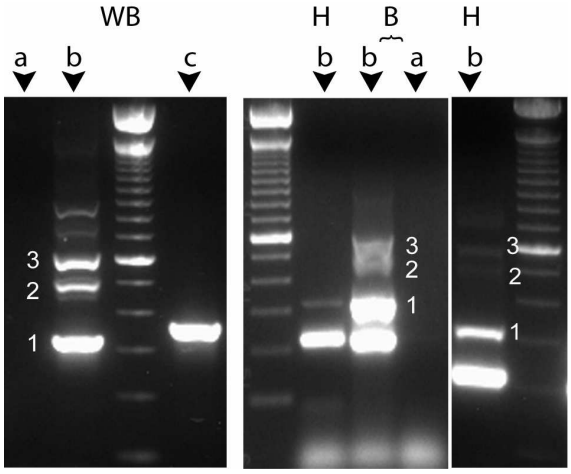


Figure 5 *S. maritima* expression of the *Sm35* region encompassing duplicated exons coding for Ig7. Whole body (WB), hemocytes (H) and brain (B). **a** negative control of b; **b** Region encompassing Ig7 coding duplicated exons of *Sm35* **c** expression of *Sm 35* constitutive exons coding for Ig9. All bands were cloned and sequenced; 1 corresponds to transcripts with exons coding for Ig6 and Ig8, missing Ig7 coding exons altogether; 2 corresponds to transcripts for which Ig7 coding exons were mutually exclusive alternatively spliced using a premature splicing site and 3 to transcripts for which Ig7 coding exons were mutually exclusive alternative splicing. The larger bands that follow correspond to transcripts with more than one Ig7 coding exon.

Alternatively spliced Dscam of myriapodes is expressed by hemocytes and nervous system

In insects Dscam diversity is used both in the nervous and immune systems. We investigated whether *Sm35* is expressed both by hemocytes and by nervous system cells of *S. maritima* by RT-PCR. The hemolymph withdrawn from two *S. maritima* individuals was rich in hemocytes (Fig. S6). To obtained nervous cells enriched tissue, the heads of three individuals were used to obtain RNA. The sequences of cloned the RT-PCR fragments shows that this gene is expressed by both hemocytes and nervous system (Fig. 5). Several different transcripts were obtained from the whole body. This result indicates that the expression of Dscam by hemocytes is not a derived character that evolved in pancrustaceans but a character that was most likely already present the ancestor of this group.

Diversity of transmembrane domains and cytoplasmic tails of Ixodes and Strigamia Dscams

We found one member of the *I. scapularis* Dscam family with two exons coding for transmembrane domains, which indicates that it might use alternative transmembrane domains through alternative splicing (Is9, Fig. S5) like the Dscam-hv of insects (Watson et al. 2005). In support of that we found one EST corresponding to the expression of Is9 where only one of the transmembrane forms is used (Fig. S5). The Dscam homologue Is13 does not contain a transmembrane domain possibly coding for a Dscam soluble form (Is13, Fig. S5). Supporting that, another EST was found in which there is no transmembrane domain, corresponding to the expression of the homologue Is13. The EST end coincides with the end of FNIII-6, i.e. the end of the ectodomains of Is13 (Fig. S5).

In *S. maritima*, the gene *Sm35* of encodes different cytoplasmic tails by alternative usage of exons (Fig. S7), indicating that this molecule might engage in different signaling pathways like the Dscam of pancrustaceans. The sequence

conservation between the cytoplasmic domains of *S. maritima* and of *I. scapularis* with the cytoplasmic tails of pancrustaceans is low (data not show). Nevertheless a few motifs are conserved and among those are motifs that belong to the so called CC0-3 motifs category in particular CC1 motifs (PTPYATT) (Prasad et al. 2007; Andrews et al. 2008) (Fig. 6).

```
Homo Robo. TTYSRPGQPTPYATTQLIQSNLSNN 124

Dugesia    NNDDEDEMLVPYATYESLSKPDSST 105
Aplysia    SFRSDEGNINPYATYNEIKPTFIPE 139
Strongyl.  EPRRHRGLADPYATFDYHDGSIYPS 126
Ixodes 6   LEGRLDYYPTPYATTRVTDIDERKL 68
       23  ECSTSAFFPAPYATTHLGTRGPEKR 72
       10  PRGDPLYFPSPYATTHISVYSGDND 69
       15  PSKDQIYYPSPYALGGREPVLHRQG 69
Stri.52294 GSHVDSDELTPYATARLADFQEHRR 61
    321807 QNSLRRGDVAPYATGHLSDHYQAEE 95
     34735 TIPRRGADPSPYATSHLTDCHHPEH 94
     Sm35  LVKGSSDEITPYATTQLPNFHYGEM 66
     24872 YTQTSLEDVCPYATYRIPESSNKAQ 98
     56727 TREGVHDDACPYATFQLSENKQNSN 102
Drosophila RHPGMEDEICPYATFHLLGFREEMD 162
DscamL2    EGNEYIEDICPYATFQLNKQTYSES 108
DscamL3    GNESEMYEISPYATFSVNGGRTGAP 92
DscamL4    KIPETSEDISPYATFQLSEAGGNMS 96
Daphnia    LYAGMDDEICPYATFHLLGFREEMD 151
DscamL2    LSDYAPDQVSPYAVFPSLTSSGGKS 104
Dscam16    DNPNQLGDITPYATFTLKPINGMDT 123
Pacifast.  LRSGGDDEICPYATFHLLGFREEMD 165
```

Figure 6 Conservation of CC1 motif PTPYATT between Human Roundabout and DSCAM family molecules from invertebrates. The numbers on the right refer to the position of the aminoacid with respect to the beginning of the transmembrane domain of the molecule. In red the CC1 motif and in blue some (relatively less) conserved flanking aminoacids. All sequences except Human Robo are from Dscam-hv or Dscam-like molecules. The comparison *Pacifastacus leniusculus*, *Daphnia pulex* and *Strigamia maritima* reveals tha the sequence GxxDEICPYATFHLLGFREEMD (underlined) is a good marker of the variable Dscam. Abbreviations; *Homo* Robo: Human roundabout; *Strongyl.*: *Strongylocentrotus purpuratus*; *Stri.*: *Strigamia maritima*; *Pacifast.*:*Pacifastacus leniusculus*.

These motifs are also present in the Dscam protein of other invertebrates but not of vertebrates. Interestingly they are also shared by vertebrate cell adhesion molecules loosely related to DSCAM such as roundabout. The comparison of several Dscam cytoplasmic tails of arthropods revealed that the residues GxxDEICPYATFHLLGFREEMD are a good predictor of Dscam genes containing domains diversified by alternative splicing (Fig. 6). Interestingly these motifs are present in Sm35 for which alternative splicing and the expression by hemocytes was demonstrated, but not in the other Dscam duplicates of *S. maritima* with several exons coding for Ig7.

Evolution of the Dscam gene family

Our data suggest that the Dscam gene with arrays of exons coding for Ig2, Ig3 and Ig7 evolved uniquely in the ancestor of pancrustaceans (Fig. 7). Nevertheless, diversification of Dscam homologues occurred in all arthropod groups either by internal duplication Ig domains or by duplications of complete genes. The genealogy of all Dscam gene reconstructions of *S. maritima* and *I. scapularis*, confirmed that the former correspond to Dscam homologues which diversified within each taxa independently. Despite their differences, arthropod's Dscams seem indeed to be more strongly related to each other than to any other homologues in the Dscam gene family forming a monophyletic group (Fig. 7). Within arthropods Dscam-Hv, Dscam-L2, of pancrustaceans form two separated clades. Noteworthy, Dscam-L2 of pancrustaceans and all the genes of *S. maritima* with Ig7 coding exon duplications, do not have a common origin.

This genealogy also demonstrates that not all insect groups share the same four Dscam paralogues.

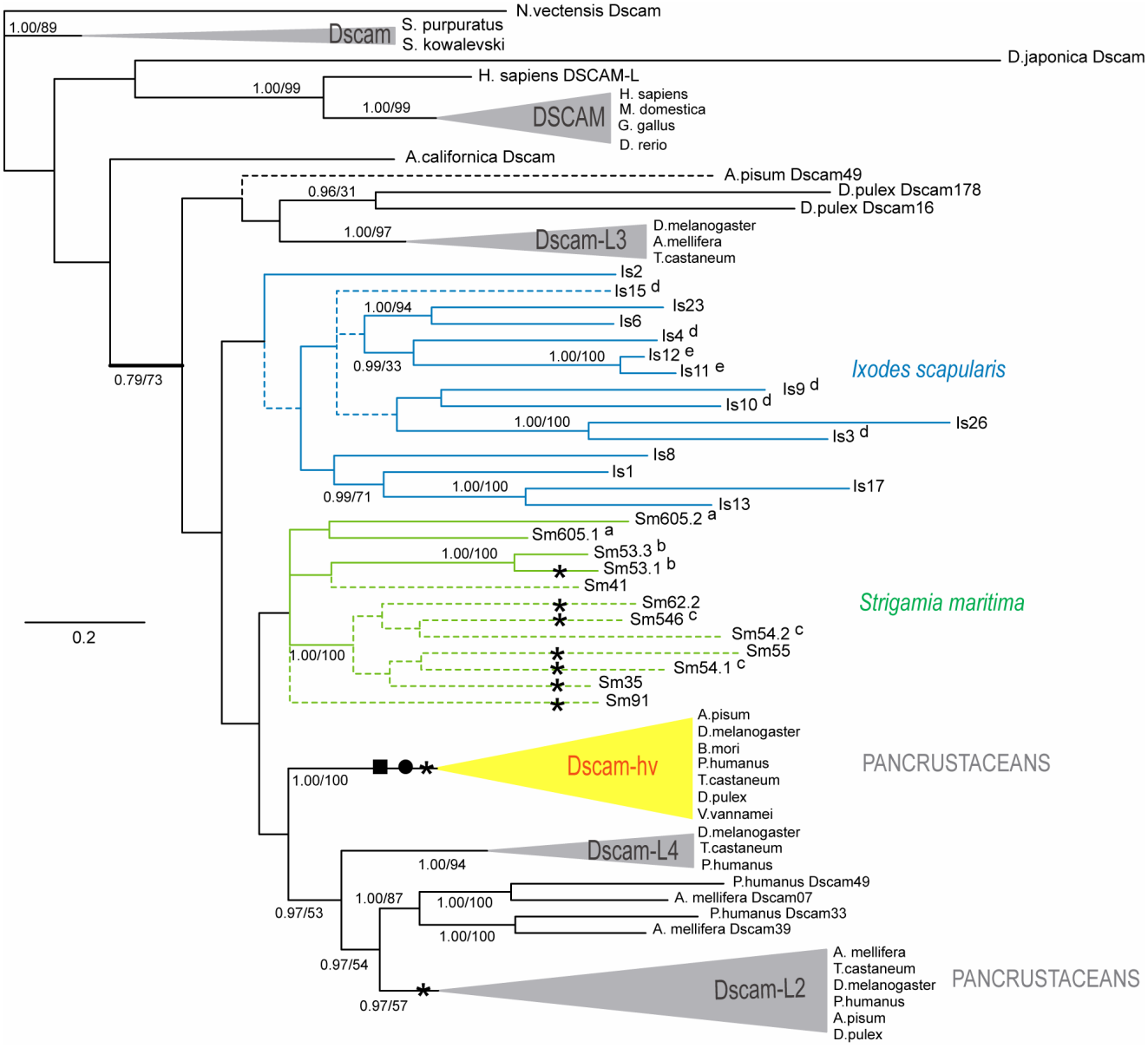


Figure 7 Maximum likelihood topology of Dscam related genes in representatives of metazoa. The tree is rooted using the Dscam sequence of the cnidarian *Nematostella vectensis*. Support values at nodes are bootstrap values relative to 1000 replicates (left value) and posterior probabilities (right value) when higher than 60% and/or than 0.95, respectively. Monophyletic clades of orthologues were collapsed for convenience. Genes with internal exon duplications coding for Ig2, Ig3 and Ig7 are indicated with ■, • and *, respectively. Genes located in the same genomic scaffold are indicated with the same superscript. The dashed branches represent incongruent branches obtained by the maximum likelihood and the Bayesian methods. The monophyletic origin of all arthropod Dscams is marked by a thicker internal branch.

Both *A. mellifera* and the lice species *P. humanus* have five Dscam paralogues. Two of them share a common ancestor and are not present in the other insect species analyzed (Fig. 7).

Contrarily to previous results (Brites et al. 2008), *Daphnia pulex* has two other paralogues besides Dscam–hv and Dscam-L2 which do not group confidently with any of the other insect Dscam paralogues.

The *S. maritima* Dscam homologues are more closely related to each other than to any other Dscam and the same is true for *I. scapularis*. In both taxa, gene duplication was followed by quick divergence such that the phylogenetic relationships among paralogues are difficult to recover (Fig. 7). The paralogues with Ig7 exon duplication do not form a monophlyletic group within the *S. maritima* paralogues.

The origins of the duplicated genes coding for Ig7

All arthropods evolved Dscam paralogues with internal exon duplications coding for Ig7. This suggests that the array of Ig7 coding exons might be the origin of the alternatively spliced exons of Dscam in arthropods. *Ixodes scapularis* and *S. maritima* Ig7 coding exons always rendered higher similarity to the Ig7 coding exons of pancrustacea than to any other Dscam in our blast searches. We tested whether a common ancestor between exons coding for Ig7 in myriapodes, chelicerates and pancrustaceans could be found, in which case we expected them to form a monophyletic groups in relation to the rest of Ig7 coding exons of other Dscams. We produced Bayesian and ML trees containing all Ig7 coding exons of all Dscam paralogous and orthologous genes of representative metazoa (Table S2), together with all Ig7 coding exons present in *Ixodes* and *Strigamia*. The results show confidently monophyletic groups of exons within species but generally low statistical confidence in the nodes that connect the Ig7 coding exons from the main arthropod groups (Table S2). This is not unexpected given that Ig7 coding exons are short sequences that, except for a few landmark amino acids, diverged extensively in the represented taxa.

The only exceptions found were monophyletic relationships between the ig7 coding exons 11.16 of *Daphnia pulex* and 9.33 of *Drosophila melanogaster* (also found by Lee et al. 2009), and between exon 7.6 of *Daphnia pulex* and 7.16 of *Apis mellifera*, indicating that these exons were probably present in the ancestors of pancrustaceans.

The alignment of all Ig7 coding exons revealed an interesting difference between all Ig7 coding exons of Dscam-hv and all the other Dscams. Between the conserved tryptophan 38 and glycine 42, all ig7 coding exons except those belonging to the Dscam-hv, have a variable nonpolar aminoacid, followed by arginine or lysine and aspartic acid (Fig. S9). This is not observed in any of the Ig7 coding exons of the selected pancrustacea species, which have a variable amino acid composition between tryptophan 38 and glycine 42, but have invariably, arginine or lysine at position 58 which was never observed outside of the Dscam-hv (Fig. S8). Curiously, exons 11.16 of *Daphnia pulex* and 9.32 and 9.33 of *Drosophila melanogaster*, for which a common origin is still

noticeable, exhibit an intermediary composition at these positions, with aspartic acid before glycine 42 and no charged amino acid at position 58. In both species, these exons are located at the end of the array. Possibly they did not diverge as much as the exons more internally located in the arrays and still retained ancestral features (Brites et al. 2008; Lee et al. 2009). According to models based in the *Drosophila melanogaster* Dscam-hv protein structure, the position 64 is at the beginning of an Ig7 domain D' strand which is involved in homophilic binding between Dscam isoforms whereas the region of Ig7 encompassing tryptophan 40 and glycine 44 has no described function (Sawaya et al. 2008). The significance of these amino acid changes is not clear, but given the prominent differences between Dscam-hv and the other Dscams they are likely to be important functionally.

Discussion

The evolution of the Dscam family

Throughout the evolution of metazoans, cell adhesion molecules (CAMs) were recruited for many different cellular functions; cell proliferation and differentiation, apoptosis, migration and parasite recognition among others (Buckley et al. 1998; Humphries and Newham 1998). Many members of this family are at least partially composed of multiple Ig domains (Chothia and Jones 1997). In some of those members, the first four Ig domains of the molecules form of a tertiary conformation called the horse-shoe structure which creates singular adhesive properties by allowing homophilic and heterophilic adhesion to similar and different proteins, respectively. The appearance of this structural feature might have allowed the expansion of a sub-family of CAMs used by nervous cells of different metazoans such as axonin, roundabout, contactin, Dscam, etc, and by immune system cells such as hemolin and Dscam. Our analysis of basal metazoan CAMs suggests that precursors of Dscam could be already present before the evolution of the Bilateria. Certain regions of the cnidarian NV_1 protein are quite conserved between *Nemastostella vectensis* and humans. Furthermore, Nv_1 shares cytoplasmic motifs with human Dscams (but not with any of the protostome Dscam homologues) denoting the usage of similar signaling pathways. This suggests that some of the Dscam features characteristic of complex groups such as vertebrates might have evolved already in early metazoans.

In vertebrates, in the flat worm *Dugesia japonica* and most likely in all other metazoans, Dscam is essential for the correct development of the nervous system (Yamakawa et al. 1998) (Fusaoka et al. 2006). The same is true for the pancrustacean Dscam-hv and Dscam-L2 which have been shown to participate in the nervous system development of *Drosophila melanogaster* (Millard et al. 2007)(Millard et al. 2007). All extant arthropod groups, pancrustaceans, myriapods and chelicerates, had extensive expansions of this gene family. This

occurred both by massive duplication of entire Dscam genes, of which chelicerates and myriapodes are an extreme example, and by extensive internal duplication of certain exons such as in Dscam-hv of pancrustaceans, and to a lesser extent in Dscam-L2 and in all the Dscam homologues of *I. scapularis* and *S. maritima* with Ig7 coding exon duplications.

In contrast to the extracellular domains of Dscam of distant taxonomic groups, homology between the cytoplasmic tails of the different metazoan Dscam cannot be traced even though certain short motifs are conserved. This suggests that evolution of the extracellular and intracellular part of the Dscam family molecules must have involved exon shuffling at different rates. The result is a number of members with highly similar extracellular domain conservation of the horseshoe distal extremity and Ig7 but with very divergent intracellular segments. That suggests that the selective pressures on the external and internal parts of the molecule in different organisms were not the same and that the properties of the receptor were accommodated to multiple signaling pathways. Additionally, alternative splicing appears to be used in many instances to diversify both extracellular and intracellular parts of the molecule. All things considered, the independent acquisition by different organisms of multiple Dscam forms, either by producing numerous protein isoforms by alternative splicing of duplicated exons or by usage of multigene families and by using different cytoplasmic tails, suggests a very strong pressure to diversify the family, mostly evident in the extant Arthropods groups analyzed.

The Dscam genes of arthropods

Despite the differences among arthropod Dscam homologues our phylogenetic analysis suggests a monophyletic origin for the Dscam family in this group. In the remaining metazoans no Dscam paralogues are known, with the exception of vertebrates in which two paralogues of Dscam (DSCAM and DSCAM-L in humans) have arisen independently of the arthropod duplicates (Brites et al. 2008). Why the evolutionary history of this gene family is so different between arthropods and the remaining metazoan groups is not easily answered. Whatever the cause may be, the genetic diversification of Dscam in arthropods has allowed the functional diversification of the gene. That is evident in pancrustaceans for which the Dscam-hv expresses diverse splicing repertoires both in nervous cells and hemocytes (the immune cells of both insects and crustaceans) (Watson et al. 2005; Dong, Taylor, and Dimopoulos 2006; Brites et al. 2008). Here we show for the first time that the expression of Dscam diversity created by mutually exclusive alternative splicing by hemocytes is not a derived character of pancrustaceans, the hemocyte cells of the myriapod *S. strigamia* also express Dscam variants created by mutually exclusive alternative splicing of Ig7 coding exons. This character was thus most likely already present in the ancestors of pancrustaceans.

It has generally been assumed that Dscam-hv evolved in all arthropods (Crayton et al. 2006; Kurtz and Armitage 2006; Lee et al. 2009; Schmucker and Chen 2009). Our data show that the Dscam gene with arrays of exons coding for Ig2, Ig3 and Ig7 evolved uniquely in the ancestor of pancrustaceans. Yet, we found a high diversity of Dscam caused by expansions of Dscam homologues in *S. maritima* and *I. scapularis*, which have occurred by several rounds of duplications of the whole Dscam gene and/or by duplication of certain Dscam domains. Furthermore, in both groups there are Dscam homologues with duplicated exons that code for Ig7 and Ig8 in the case of *I. scapularis*. Interestingly, the gene expansions in both taxa seem to have occurred independently given that Dscam homologues are always more related within than between those taxa. A striking aspect of these gene expansions is that they reveal a highly dynamic interaction between Dscam paralogs through which many kinds of genetic arrangements were possible. Furthermore, a large part of the genes found in *I. scapularis* and in *S. maritima* seems to be functional given that only some pseudo-exons (exons with incorrect splicing sites or shifts in reading frame) were observed. In addition we show that duplicated exons coding for Ig7 in *S. maritima* can be mutually exclusive alternatively spliced, adding isoform diversity to the diversity created by the expression of the numerous whole duplicated genes.

In both *I. scapularis* and *S. maritima* there are Dscam molecules with signaling capacities similar to Dscam-hv. An interesting characteristic of the transmembrane domains of both groups Dscams is that they are unusually rich in cysteines (Table S3). Cysteines are important binding residues that could favour the formation of complex membrane-bound Dscam multimers or associations of Dscam with other proteins. This feature might allow those Dscam members of *Ixodes* and *Strigamia* to be engaged in different cellular functions. The cytoplasmic tails of several Dscam members in both *Strigamia* and *Ixodes* contain furthermore a number of motifs common to the Dscam-hv of pancrustaceans (Brites et al. 2008), namely numerous SH2 binding sites (Schmucker et al. 2000), endocytosis/phagocytosis motifs (Indik et al. 1995) and several immunoreceptor tyrosine-based inhibition and immunoreceptor tyrosine-based activation motifs, ITIMs and ITAMs, respectively (Barrow and Trowsdale 2006; Daeron et al. 2008) (Table S3). This indicates that these Dscam genes can have similarities to Dscam-hv in their signaling capacities and protein associations. We have found that CC1 motifs (PYATT) (Prasad et al. 2007; Andrews et al. 2008) present in all arthropod Dscams and in the Dscam proteins of other invertebrates but not of vertebrates. Interestingly they are also shared by vertebrate CAMs loosely related to Dscam such as roundabout. In roundabout molecules, these motifs can be involved in axon guidance signaling pathways and importantly, in leukocyte mobility control via heterologous binding with the ligand SLIT (Prasad et al. 2007). The latter function could indeed be shared with arthropods

given the expression of Dscam by hemocytes. The homophilic binding between Dscam isoforms plays an important role in axon guidance (Matthews et al. 2007; Meijers et al. 2007; Wojtowicz et al. 2007) but heterologous binding to the ligand Netrin, has been demonstrated to contribute also to axon guidance both in *Drosophila* and in mammals (Andrews et al. 2008). In sum, these aspects suggest that the expression of Dscam diversity by arthropod hemocytes could be related to hemocyte mobility which in turn could have consequences both for immunity and organogenesis.

The diversity of Dscams found in those animals recapitulates the Dscam-hv of pancrustacea, i.e high diversity of Dscam ectodomains, Dscam molecules with mutually exclusive alternative splicing of internal duplications, Dscam molecules with alternative transmembrane domains such as in insects, Dscam soluble forms like in pancrustacenas (in decapode crustaceans a Dscam soluble form is encoded in the genome whereas in insects is produced by proteolytic cleavage of membrane bound forms (Chou et al. 2009) (Schmucker et al. 2000). The fact that different groups of pancrustaceans have different Dscam paralogues (Fig. 7) suggests that their most recent common ancestor had large diversity of Dscam genes, similarly to *S. maritima* and *I. scapularis*, from which different paralogues were retained in the extant pancrustacean groups. We speculate that extensive Dscam duplications, gene rearrangements and the mutually exclusive alternative splicing mechanism found for Ig7 coding exons seen in *Ixodes* and *Strigamia* were the raw material from which Dscam-hv evolved in the ancestors of the pancrustaceans.

The origin of Dscam-hv in pancrustaceans

Some duplications of Dscam homologues in *Ixodes* and *Strigamia* occurred within short genomic regions as demonstrated by the fact that a number of contiguous genes are more similar to each other than to other genes (i. e. Fig. 7, *Is12* and *Is11*; *Sm53.1* and *Sm53.3; Sm605.1* and *Sm605.3*). Other duplications are found in different genomic scaffolds indicating that they occurred over longer regions in the genome (i.e. Fig. 7, *Is26* and *Is3*) and genes such as *Is53* are chimeras between other duplicated genes (Fig. 3). This situation could have arisen due to mispairing (Zhang 2003) during meiotic homologous recombination, a common mechanism of duplication and the likely mechanism underlying the duplications in Dscam-hv arrays of exons. We propose that a similar mechanism created a large number of Dscam duplicates in the ancestor of pancrustaceans, and is at the origin of the arrays of alternative duplicated exons that confer diversity to half of Ig2 and Ig3 domains and to the complete Ig7 domain of extant pancrustacea. The intriguing question is why only those exons duplicated and not others. Structural aspects of Dscam-hv and the molecular basis of its role in the nervous system, provide insights into how this might have been achieved.

An important basis for the molecular action of Dscam is the formation of Dscam dimers trough homophilic binding of identical Dscam isoforms, leading to a self-avoidance behavior of nervous cells essential for neural wiring in *Drosophila melanogaster* (Hughes et al. 2007; Matthews et al. 2007, Soba et al. 2007; Wojtowicz et al. 2007). Remarkably, the Dscam regions involved in dimer formation are fractions of Ig2, Ig3 and Ig7 domains coded by the duplicated exons (Meijers et al. 2007; Sawaya et al. 2008). In this way the genetic diversification caused by the duplications, coupled with the strong specificity of Dscam's homophilic binding, provide a huge repertoire of highly specific "key-locks" which nervous cells exploit extensively (Hughes et al. 2007; Matthews et al. 2007; Meijers et al. 2007; Soba et al. 2007; Wojtowicz et al. 2007; Sawaya et al. 2008). We propose that the homophilic binding between Dscam molecules having internal duplications coding for Ig2, Ig3 and Ig7 was the mechanism that drove selection on all duplications that coded for those domains because that increased the number of possible Dscam dimers, providing cells with a diverse self non-self recognition system. In this way duplications that conferred direct functional diversity would be selected whereas others would be lost by drift or by purifying selection. We speculate that internal duplications coding for other Ig domains might have occurred (as the Ig8 duplications of *I. scapularis* suggest), but only the ones participating in half of Ig2, half of Ig3 and Ig7 domains have been selected based on structural and functional features of Dscam in the pancrustacea ancestors.

Another possible explanation is that the regions coding for half of Ig2 and Ig3 and the complete Ig7 could be more prone to duplication (like suggested by the apparent independent duplications coding for Ig7 and Ig8 in *Is27*, *Is28* and *Is29* genes), maybe because they reside on recombination hot spots. A third possibility still, suggested by the existence in *Strigamia* and *Ixodes* of contiguous Dscam genes separated in some case by relatively short genomic sequences, is that the transcription of such contiguous genes is not totally independent. This could produce a step-wise expression of these genes similar to alternative splicing. Under this scenario, again based on the selection imposed by the specificity acquired via dimers formation, the composition of the ectodomains of the molecule like it exists in extant pancrustacea could have been shaped mainly by domain lost.

The origin of the mutually exclusive alternative splicing of the duplicated exons

The extraordinary molecular diversity of Dscam-hv expressed by nervous cells and by the hemocytes of pancrustaceans is achieved via a process of mutually exclusive alternative splicing of the internal exon duplications coding for half of Ig2 and Ig3 domains and the complete Ig7. This process ensures that only one exon per array of duplications is present in the mature RNA. Throughout evolution alternatively spliced exons appeared as a transition from constitutive

to alternative exons among other mechanisms (Ast 2004). The Ig2 and Ig3 exon duplications encode only half domains, thus any duplicated exons transcribed constitutively would render a non-functional protein and be deleterious. In the case of Ig7, given that it is encoded by a complete exon, exon duplications constitutively expressed would potentially code for a functional protein with several Ig7 domains. A plausible scenario is that the regulators of the alternative splicing mechanism of Ig7 were used in the ancestors of the pancrustaceans to splice exon duplications coding for Ig2 and ig3 domains. In that case we would predict that the three arrays of duplications have in pancrustaceans at least some common regulating features.

We could not show that the duplicated alternatively spliced exons coding for Ig7 in *S. maritima* and in the pancrustacean Dscam-hv have a common origin due to the little phylogenetic signal present in such short region which diverged extensively among such distant taxonomic groups.

Whatever the case may be, there was convergent evolution in different arthropod groups to generate Dscam diversity. The reasons why this diversity was selected for are probably related to the self vs non-self cell recognition system created by the specificity of binding between different Dscam molecules. Interestingly, exon duplicates of Dscam-hv in pancrustaceans seem to have diverged mainly under neutral evolution (Brites et al. 2011), suggesting an evolutionary scenario in which accumulating aminoacid diversity was more important than the exact aminoacid sequences created.

Aknowledgements

We are very thankful to Michael Akam for providing us with privileged access to the genomes of *Strigamia maritima*.

REFERENCES

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104-2105.

Agarwala, K. L., G. Subramaniam, Y. Tsutsumi, T. Suzuki, A. Kenji, and K. Yamakawa. 2001. Cloning und Functional Characterization of DSCAML1, a Novel DSCAM-like Cell Adhesion Molecule that Mediates Homophilic Intercellular Adhesion. Biochem Bioph Res Co:760-772.

Andrews, G. L., S. Tanglao, W. T. Farmer, S. Morin, S. Brotman, M. A. Berberoglu, H. Price, G. C. Fernandez, G. S. Mastick, F. Charron, and T. Kidd. 2008. Dscam guides embryonic axons by Netrin-dependent and -independent functions. Development **135**:3839-3848.

Ast, G. 2004. How did alternative splicing evolve? Nature Reviews Genetics **5**:773-782.

Barrow, A. D., and J. Trowsdale. 2006. You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signalling. Eur J Immunol **36**:1646-1653.

Brites, D., F. Encinas-Viso, D. Ebert, L. Du Pasquier, and C. R. Haag. 2011. Population genetics of duplicated alternatively spliced exons of Dscam in *Daphnia* and *Drosophila*. *PloS ONE* **6**:e27 947.doic:910.1371/journal.pone.0027947.

Brites, D., S. McTaggart, K. Morris, J. Anderson, K. Thomas, I. Colson, T. Fabbro, T. J. Little, D. Ebert, and L. Du Pasquier. 2008. The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Molecular Biology and Evolution **25**:1429-1439.

Buckley, C. D., G. E. Rainger, P. F. Bradfield, G. B. Nash, and D. L. Simmons.

1998. Cell adhesion: more than just glue (Review). Molecular Membrane Biology **15**:167-176.

Budd, G. E., and M. J. Telford. 2009. The origin and evolution of arthropods. Nature **457**:812-817.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution **17**:540-552.

Chothia, C., and E. Y. Jones. 1997. The molecular structure of cell adhesion molecules. Annual Review of Biochemistry **66**:823-862.

Chou, P. H., H. S. Chang, I. T. Chen, H. Y. Lin, Y. M. Chen, H. L. Yang, and K. C. H. C. Wang. 2009. The putative invertebrate adaptive immune protein Litopenaeus vannamei Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. Developmental and Comparative Immunology **33**:1258-1267.

Crayton, M. E., 3rd, B. C. Powell, T. J. Vision, and M. C. Giddings. 2006. Tracking the evolution of alternatively spliced exons within the Dscam family. BMC Evol Biol **6**:16.

Daeron, M., S. Jaeger, L. Du Pasquier, and E. Vivier. 2008. Immunoreceptor tyrosine-based inhibition motifs: a quest in the past and future. Immunological Reviews **224**:11-43.

Dong, Y., H. E. Taylor, and G. Dimopoulos. 2006. AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biol **4**:e229-.

Drummond, A., and K. Strimmer. 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics **17**:662-663.

Fusaoka, E., T. Inoue, K. Mineta, K. Agata, and K. Takeuchi. 2006. Structure and function of primitive immunoglobulin superfamily neural cell adhesion molecules: a lesson from studies on planarian. Genes to Cells **11**:541-555.

Graveley, B. R. 2005. Mutually exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell **123**:65-73.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52**:696-704.

Hattori, D., S. S. Millard, W. M. Wojtowicz, and S. L. Zipursky. 2008. Dscam-Mediated Cell Recognition Regulates Neural Circuit Formation. Annual Review of Cell and Developmental Biology **24**:597-620.

Hughes, M. E., R. Bortnick, A. Tsubouchi, P. Baumer, M. Kondo, T. Uemura, and D. Schmucker. 2007. Homophilic Dscam interactions control complex dendrite morphogenesis. Neuron **54**:417-427.

Humphries, M. J., and P. Newham. 1998. The structure of cell-adhesion molecules. Trends in Cell Biology **8**:78-83.

Indik, Z. K., J. G. Park, S. Hunter, and A. D. Schreiber. 1995. The Molecular Dissection of Fc-Gamma Receptor-Mediated Phagocytosis. Blood **86**:4389-4399.

Kreahling, J. M., and B. Graveley. 2005. The iStem, a Long- Range RNA Seconday Structure Element Required for Efficient Exon Inclusion in the *Drosophila Dscam* Pre-mRNA. Molecular and Celular Biology **25**:10251-10260.

Kurtz, J., and S. A. Armitage. 2006. Alternative adaptive immunity in invertebrates. Trends Immunol **27**:493-496.

Lee, C., N. Kim, M. Roy, and B. R. Graveley. 2009. Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. Rna **16**:91-105.

Lefranc, M.-P., and G. Lefranc. 2001. The Immunoglobulin Facts Book. Academic Press, London.

Letunic, I., T. Doerks, and P. Bork. 2009. SMART 6: recent updates and new developments. Nucleic Acids Research **37**:D229-D232.

Matthews, B. J., M. E. Kim, J. J. Flanagan, D. Hattori, J. C. Clemens, S. L. Zipursky, and W. B. Grueber. 2007. Dendrite self-avoidance is controlled by Dscam. Cell **129**:593-604.

Meijers, R., R. Puettmann-Holgado, G. Skiniotis, J.-h. Liu, T. Walz, J.-h. Wang, and D. Schmucker. 2007. Structural basis of Dscam isoform specificity. Nature **449**:487-491.

Millard, S. S., J. J. Flanagan, K. S. Pappu, W. Wu, and S. L. Zipursky. 2007. Dscam2 mediates axonal tiling in the Drosophila visual system. Nature **447**:720-U714.

Miller, M., M. Holder, R. Vos, P. Midford, T. Liebowitz, L. Chan, P. Hoover, and

T. Warnow. 2009. The CIPRES Portals. CIPRES.

Nylander, J. A. A., J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics **24**:581-583.

Olson, S., M. Blanchette, J. Park, Y. Savva, G. W. Yeo, J. M. Yeakley, D. C. Rio, and B. R. Graveley. 2007. A regulator of Dscam mutually exclusive splicing fidelity. Nature Structural & Molecular Biology **14**:1134-1140.

Prasad, A., Z. Qamri, J. Wu, and R. K. Ganju. 2007. Pivotal advance: Slit-2/Robo-1 modulates the CXCL12/CXCR4-induced chemotaxis of T cells. Journal of Leukocyte Biology **82**:465-476.

Sawaya, M. R., W. M. Wojtowicz, I. Andre, B. Qian, W. Wu, D. Baker, D. Eisenberg, and S. L. Zipursky. 2008. A double S shape provides the structural basis for the extraordinary binding specificity of Dscam isoforms. Cell **134**:1007-1018.

Schmucker, D., and B. Chen. 2009. Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. Genes Dev **23**:147-156.

Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. l. Zypursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity Cell **101**:671-684.

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. Proceedings of the National Academy of Sciences of the United States of America **95**:5857-5864.

Shapiro, L., J. Love, and D. R. Colman. 2007. Adhesion molecules in the nervous system: Structural insights into function and diversity. Annual Review of Neuroscience **30**:451-474.

Soba, P., S. Zhu, K. Emoto, S. Younger, S. J. Yang, H. H. Yu, T. Lee, L. Y. Jan, and Y. N. Jan. 2007. Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization. Neuron **54**:403-416.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology **56**:564-577.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics **25**:1189-1191.

Watson, L. F., F. T. Püttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel, and D. Schmucker. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects Science **309**:1874-1878

Watthanasurorot, A., P. Jiravanichpaisal, H. P. Liu, I. Soderhall, and K. Soderhall. 2011. Bacteria-Induced Dscam Isoforms of the Crustacean, Pacifastacus leniusculus. Plos Pathogens **7**.

Wojtowicz, W. M., W. Wu, I. Andre, B. Qian, D. Baker, and S. L. Zipursky. 2007. A vast repertoire of Dscam binding specificities arises from modular interactions of variable ig domains. Cell **130**:1134-1145.

Yamakawa, K., Y.-K. Huo, M. A. Haendel, R. Hubert, X.-N. Chen, G. E. Lyons, and J. R. Korenberg. 1998. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. Hum Mol Genet **7**:227-237.

Yu, H. H., J. S. Yang, J. Wang, Y. Huang, and T. Lee. 2009. Endodoamin Diversity in the Drosophila Dscam and Its Roles in Neuronal Morphogenesis. Journal of Neuroscience **29**:1904-1914.

Zhang, J. Z. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution **18**:292-298.

SUPLEMMENTARY MATERIAL

Table S1- Accession numbers of Dscam homologues and other CAM proteins from selected metazoan representatives.

Species		Gene accession number
Homo sapiens	Human	DSCAM-L aal57166.1
		DSCAM aac17967.1 NCAM X16841
		L1CAM NM_024003
		Roundabout3 AK056544.1
		Roundabout4 AK289769.1
		Axonin AB587327.1
Gallus gallus	chicken	XM_416734.3
Danio rerio	Zebra fish	aat36313.1
Monodelphis domestica	Opossum	XM_001370616
Manduca sexta	tobacco hornworm	Hemolin MOTP4A
Drosophila melanogaster	Fruit fly	Dscam-hv AF260530 Dscam-L2 cg42256 Dscam-L3 cg31190 Dscam-L4 cg42330 Roundabout
Apis mellifera	Honey bee	Dscam-hv AAT96374 Dscam-L2 BAF03050.1
		Dscam07 XM_392207
		Dscam-L3 XM_396307
		Dscam39 XM_392224.4
Tribolium castaneum	Flour beetle	Dscam-hv NP_001107841.1 Dscam-L2 XP_967655.2
		XM_963226
		XM_967798
Acyrthosiphon_pisum	Pea aphid	XM_001951649
		XM_001949227
		XM_001950975
Pediculus humanus		XP_002432838.1
		XP_002423033.1
		XP_002424921.1
		XP_002432149.1
		XP_002429302.1
Litopenaeus vannamei	Whiteleg schrimp	Dscam-hv GQ154653
Daphnia pulex	Water flea	Dscam-hv EU307884
		Fleabase scaffold 6
		Fleabase scaffold 16
		Fleabase scaffold 178

Bombyx mori	Silk moth	Dscam-hv
Strongylocentrotus purpuratus	Sea urchin	Dscam Xp793690
Saccoglossus kowalevskii	Acorn worm	XM_002742216 mRNA
		Axonin NM_001168034.1
Aplysia californica	Sea slug	Dscam ABS30432.1 mRNA
Dugesia japonica	Flatworm	Dscam Ab249988
Nematostella vectensis	Starlet sea anemone	Dscam like JGI scaffold_239
Amphimedon queenslandica	Demosponge	Dscam like http://reefedge.sols.uq.edu.au/genome/blast/blast_link.cgi

Figure S1 Bayesian topology of a partial region (Ig8 to FNII-2) of Dscam related genes in representatives of metazoa. The tree is rooted using the Dscam sequence of the demosponge *Amphimedon queenslandica*. The nodes' support values depicted are posterior probabilities when smaller than 0.95. Genes located in the same contig are presented in the same color.

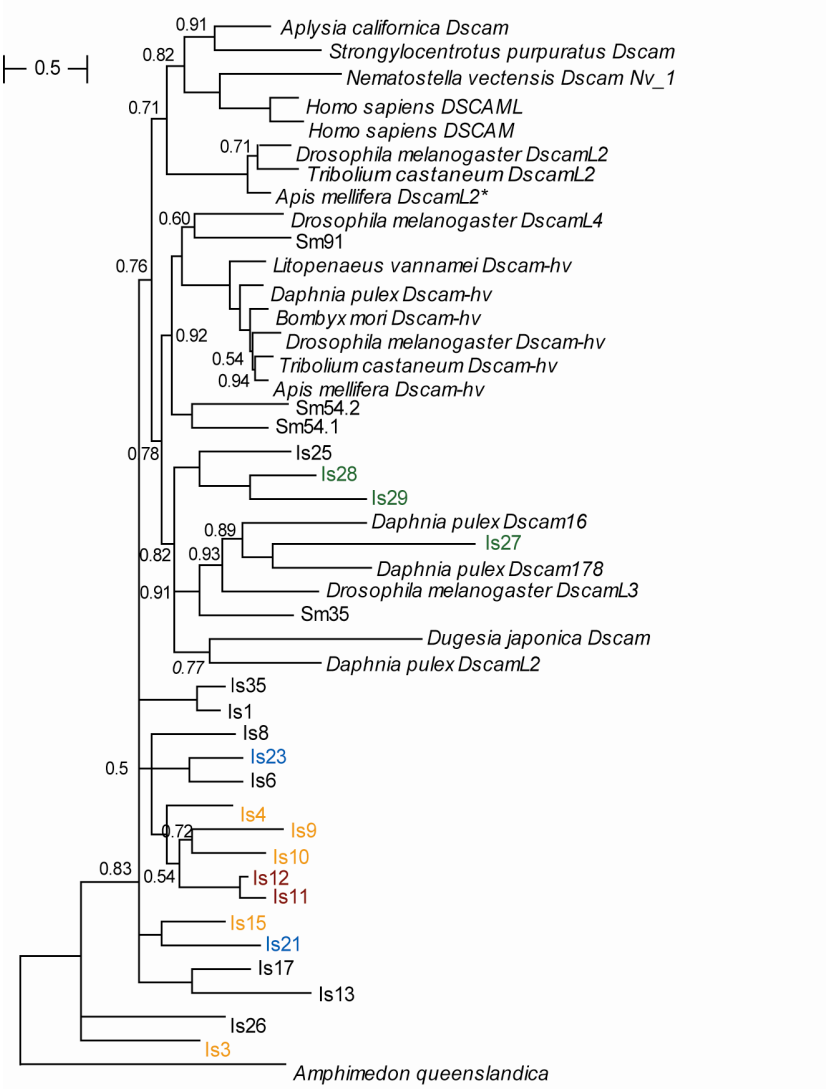
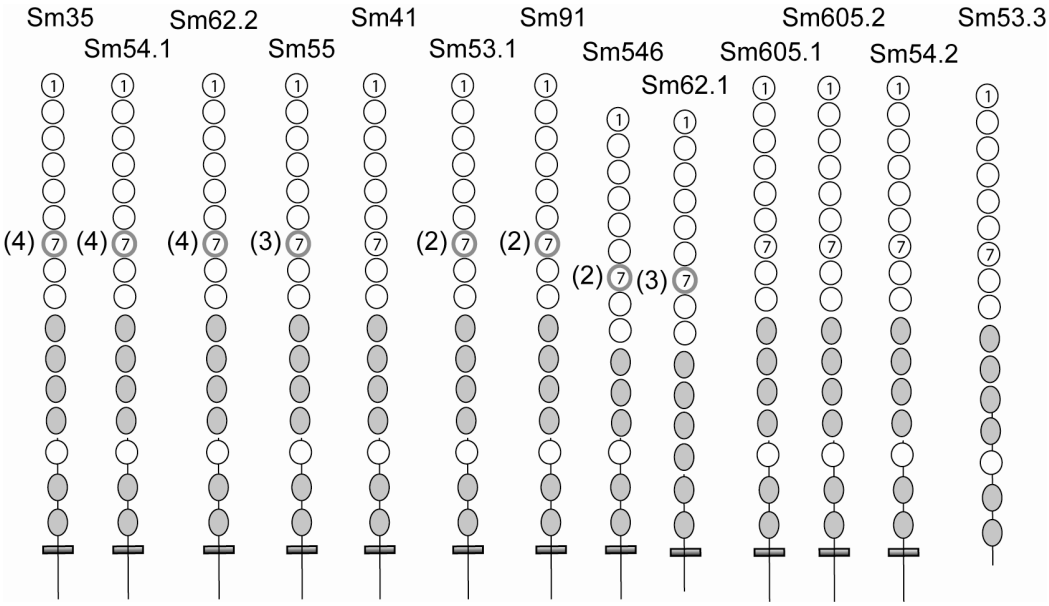


Figure S2 – A) *Strigamia maritima* reconstructions of Dscam homologues. The round circles represent Ig domains whereas the grey ellipses represent FNIII domains. The Ig7 domains which are coded by several possible exon are represented in bold and the number of possible exons is indicated in brackets. **B)** Aminoacid sequences of the *S. maritima* reconstructions. The genomic scaffold containing the gene reconstructions id indicated at the top of each reconstruction. The underlined sequences form the transmembrane domains. The domain homology of the predicted sequences when uncertain is followed by ?. In the case of *Sm35* the leader of the molecule and the regions comprising Ig7 to the transmembrane domains were confirmed by RT-PCR and the cytoplasmic domains were obtained by EST analysis. All other members were at least partially confirmed by analysis of transcripts.



B)

scf7180001248546

Sm35

Leader
LRGRSECARRRTMDTFFNLTLFLTVFCQLFL
Ig1
FAQTPTVTTPEQSIEFLQEPPDLVDFSNSTGTRIVCAASGSPTPTISWLVSDGNQVTNVTSLRQVNLDGTLVFPPFRAED
YRQDVHAVVYKCVASNVIGTIISRDVNVR
Ig2
VLQPYDVYVYDVYVIKGNTAVFRCHVPSFLVDYVKVTSWVRDSAFVIQSTFADVTSYHFSLFYQQDGKYIVMPTGELYVR
DVAANDAMTTFRCQTQHRLTGEVKMSATAGRLFVT
Ig3
VTEPQGKVQPRVTDSKTSIKANQHDTVVLPCIAQGHPVPAPKWFTKVANGHLLPVYVGDRIHQPNGALVIRDAEVADTGT
YVCVISNNASSERIETSVAIT
Ig4
VEVQPSTLLGELGKSATFRCHVSSFPISSLYWLKDGRPLPMPGLSLPTTETVLVESVRLTDRGMFQCMAKRGFESAQGTA
ELKIG
Ig5

IIPAFRAVFEERVLQPGPSLTLQCLTYGSPKPQVSWLVEGMILPEGNERSSVRDHVDATGNVVSRLTVSKVRPEDGGVYK
CISTNLAGTIEHFTRINIY
Ig6
RPGVRSRPKMTAVAGDNVMLTCPMYGYPIDLITWEKGVILPINLRQTVLPNGTLVIEKIQRATDSGKYTCIVQNKQGQSA
RGDVEVIVM
Ig7.1
VPPKITPFSFQEELLREGMRARLQCVVSEGDLPVTIKWFKDGRVIPAELGVVVRELDDVSSILAIGSVAPRHNGNYTCVA
TNDAASASHTAALFVN
Ig7.2
VGPKIIPFAFLDDQFYKGMRAHVTCAVSQGDLPITFSWSKDGWEIPPSMGVLTRSYDQHASSLTIENVSSEHTGNYSCEA
SNEAAIVQYTASLLVH
Ig7.3
VAPKIIPFSFQDEHLFEGVLARISCVVYQGDLPLTILWMKDGRPISPDLGITRRDIDDYSSILTIEKVQTTHNGNYTCVV
SNDAATVNYTAQLTVY
Ig7.4
VPPKIVPFSFQDEHLFEGMLVRVSCVISRGDLPLSITWEKDGIPIRQAPGIMVRAFDEYSSILSIDPVLPRHSGNYSCIA
HNAAGSASFTTQLLVN
Ig8
VPPRWIIEPLDSTAVKGETAMLHCKADGFPPPEISWMKTEGSSPNAERKPIISNYDTEVLYNGTLLIRQAEESSDGYYFC
RAANGVGEGLSKVVRVMIH
Ig9
VPTHFELRFSNHSTHRGEDARLKCEASGDLPIAITWRFTGESIDQRVDSRYKITETTSENGVHSEFMIHNTERKDTGMYS
CLGANKFGSDEIKLQLVVQ
FNIII-1
EAPDPPKITKLESVGNRSVHLTWSEPFDGNSKLIKYLIQYKPKSASWDNELLAPNITLDGTKLKAIVRNLFPATTYHFRL
FAENIVGTSLSSDIGTVDIEEE
FNIII-2
PGTPPRDVDCEASDPQTLRVTWKSPEKDHYTGNIRGYYIGYKIYNSTDPYNYHSVEVPDDYAEDLVFRITDLRMYAQYSV
IVQAYNDRGRGPNSPELLVMTSED
FNIII-3
PSASPTDVSCSVLTSQTINVNWQVLSLHAVNGMLRGYKVLFKPADEWDTTINQKTTDTNKLTLRNLEKSVNYSIQVLAFT
RVGDGPKSDPVYCKTYE
FNIII-4
VPGPPAQIKAIPTSLDSILVAWKPPTRPNGVIIRYNVYIRDAADTHNILGSEIHSNRDSIPRSEEYLDATKFTQNGDVTT
HEIKGLKKNRRYEFWVTAATTVGEGQSTQVIAQSPLGP
Ig10
VGATAASFSDIITSPWKHEIRLPCLAVGTPLPQRKWISGRIVKANRKVRILVDGTLVLKDIDHGDAGNYTCMVQNKIGED
KITYTLII
FNIII-5
VPPSTPTLTVVSTSLTAIELQWKPEPEETTPISGFILHYKREFGQWETINLKSDQLSFRLENLWCGTKYLVYVQGYNKIG
VGTASEIITATTEGS
FNIII-6
VPEVPNKEILLTEGPTFVTITLDGWPLTGCPIMYFVVEYKQLQTSQWVLVSNNAKAEQKKVVVPGLNPGTWYTVKVTAHS
SAGSTIAKYNFATLTAEG
GTVGPEVIIEIQEEKGVLFYLDLRVIIPIAISLLILFVVLLIMCIYFRRRNHDDRFVK
Cytoplasmic Tail
Exon1
GRIINGIVKSSSKFSSSSSTVKNYSVDSFGNGQRSTESIARRYPSISDKLVK
Exon2
EIVCSWNLGLFVSKSQDCKMKYCTWEVCFFLILSRIMSRTKKAILLDSFIIHRDSPRSRSAP
Exon3
GSSDEITPYATTQLPNFHYGEMKTFGERKSGASPFSGGG
Exon4
SDNEENLIQNTNTQKRVKKQSGEQIARPKSDGAVV
Exon5
AAAYPRPEPDGKAAWATGQPERGFSSQTGFPVGQS
Exon6
AARLPDSSMTRANSGGPSPRQQASPGDTKWRIVQRNLGNISKAKVHGV
Exon7
SSSGTQETTFIFPRTPDEVGVTPTMMSSDPTERYDEPILPPS
Exon8
AFQNKGKTDQTQADPTEGSKLLK
Exon9
SLVSCK

scf7180001248648 – 3 Dscam duplicates

Sm546

Ig1
NEPPRLVEFSNNTGAKIECTATGDPTPKVTWSLSDGTSVTNIASLRQVHADGTLVFPPFSASDFRQDIHAAVYRCVASNA
VGVVVSGDVQVK
Ig2
VLLQPYIVHIYDVYAILGNTAVMKCHVPTFLLDYVHVTSWIRDAAFVIQTTADGKYVILPSGELHIREVNPKDAMTNFRC
QTHHTLTGETRLSASAG
Ig3
EPQGNVSPRILTTQTAVHVRQGEAAILSCVAQGYPVPNTVKGSKGDLQLLRLGDRVSQKVDALVIRGARVSDSGTYVCVA
NNLVG
Ig4
APLHAKIEPAVLVAEIKKPAAFACVISGSPVSSVTWMKDGKPIVSPKPVRAAYNEKLRIESVTTEDRGMYQCIVENDYQI
RQATAELRLG
Ig5
DIAPEFLSVFEEKLQQPGTSVSLRCVARGVPLPQITWFLDDLPLPRSDRFHTDTYINRKGDRVSILNVTHMRVEDGGVYK
CESQSSAGVVQHFARINIY
Ig6
PAVRPMPKMSVVAGHDVRINCAMYGYPIESVDWEKGAAIPLDLRRMMLANGTLLIGSVERSTDSGRYRCSVRNKQGNTGT
GEVEVVV
Ig7.2
VSPKIIPFSFQDEYLREGTQARVMCALIEGDPPVKFQWLKDSRPIPSAGMAGIMVRNFDDFTSILTISNVASHHRGNYTC
VAENAAASAAHTTPLKVN
Ig7.2
VPPTILPFSFQDEHLLEGMLASVSCVVSRGDLPLSLSWEKDGLPLVPSAAKGVNIMAHGDSMSILSIGPAFPVHNGNYTC
VASNVASTMRYTAHLSVK
Ig8
VPPRWTLEPKNTMVLFGRSTVIHCQAEGFPPPSITWMKARGTEVTDQTDVLESGDEFHVFQNGSLLVKHATESHRGYYFC
AATNGIGTGLSRGVFLQVH
Ig9
VPAEIETKMQSFTVVEGEIIRARCEAKGDHPIDFTWSTDGQTIESGQHSYYFKDHVTPSRAVSELTVTNAQKMDTRIFVC
MARNPYGGDMANIQIIVQ
FNIII-1
IPDAPKIIKIADNGNRSVELAWNPPYDGNSRITKYIVQYKPIQWTDEVSNLSVSGGQSSVIIRGLTPSVGYHFRMFSENT
VGLSPPSDVVSVTMEDE
FNIII-2
APGAPPQDVEVEAMDPQTLRVMWKPPPKEMWNGAIRGYYVGFRISGTEDPFNVQTVEVPDEYMEEMKLRIPDLQKYTQYG
VMVQAFNDKGLG
FNIII-3
APPAEIKVLPLSTETVLVVWKPPSRPNGYIIKYNVYARELEDENLRSETHSSRDAMQRTSQHNFPTKHSVRGDAVQYEVK
GLKANQRYEFWVTATT
Ig10
VISIFFIVFSWLISYTAAPKSASFNEGVITAWKKEVNLTCRGVGQPSPDREWSFSSGRISYANPDGSLVIKNAQLADVGN
YSCRLFNRNGEDY
FNIII-5
PPSAPIVRVLSTTLTSIELRWSAETSERSPLQYMLHYQSDNGQWETKEIDSDYERYRLENLLCGTKYNIFVEAYNKIGLL
SEPCETIITFTEG
FNIII-6
APTRPPKDRLIDEGIGSITLHLDSWTTNGCPILYFTVEYKIRDTDEWISVEGGAKNTDKKNFLLDELDAETWYNVRMKAY
TSAGITEGIYTVGTLTLSG
IIIPVVVVLIILFLSITTICFVI
Cytoplasmic Tail
NRKRAEDKRTKGKTPLDVKNAASALGLPSGKEFNMNHAQNSDQLSRRYTSTPTRVALVFNNLRSTLKRKKDV
KRAIHPCENLTDDGSYLLLITSG

Sm54.1 Transcripts_eggs_Locus_46219

Ig1
EPSNHVEFSNETGVSINCTAHGIPEPLVTWVRTDGSLVSTVPSLRRVMADGSLVFPPFNADEYRAEVHTATYKCMASNKL
GTVVSRDVNVKA
Ig2
IMDTKDVVRVKQGETAVLPCVTQGIPVPVTTWFGKVHHEQVLPLHVGARLQQTSGALIISDTRLADSSAYICVANNTGGT
DRAETALTVTVPLS
Ig3
VKVQPPHTVADVGSSVTFTCEVTGIATSFSWLKNGRPIRVDRVRPVTADTVRIDSVQPEDRGMYQCMARNGLESAQGAAE
LRLGDK
Ig4
VPLSVKVQPPHTVADVGSSVTFTCEVTGIATSFSWLKNGRPIRVDRVRPVTADTVRIDSVQPEDRGMYQCMARNGLESAQ
GAAELRLGG
Ig5
SKPDFRETFAEKIEYPGGFVSLPCVATGSPPPHFKWTLDGIVVAEDERVSTSEMSDENGNVVTSLNITDLVVEDGGLYSC
HAINRIGSTEHGARLHIYGR
Ig6
PVVRSHLKLSAVAGERSIINCPSYGYPIEKYSWEKDGVSLPDNIRQTVYVNGTLVISEVRKVYDSGRYTCIIRNNDHIAR
GDVEIVVL
Ig7.1
VPPKIAPFSFQEELLREGMRARLQCVVSEGDLPLSIKWVKDGGDVPPTLGVLIRDLDEFSSILTINSVTPRHNGNYTCVV
ANHVATIHSTAELYVN
Ig7.2
VPPKIIPFNFIDDQFYMGMRAHITCAVSQGDLPIAFQWLKDGSEISPTLGVATRYYDQHANSLSIESVTSKHSGNYTCIA
HNVAGTAMHSAQLLVH
Ig7.3
VAPKIIPFSFQDDHLFEGVLAQISCVVYQGDLPLEIDWLKDGIPVAADTGLTLRQIDDYSSVLTIGSVQRKHSGNYTCVA
SNSAASSNFSASLTVN
Ig7.4
VPPKIVPFSFQDDHLFEGMLVRVSCVVSRGDLPLTIRWTKDGALIPPSLGVTLRDFDEYSSVLSIESVAIVHNGNYTCYA
NNSAGKASHTAQLLVN
Ig8
VPPRWRIQPKDSSVLLGREVLLNCQADGFPKPKIKWMKAEGGNIIQHRDVIHMSNVHVLSNGSLHIKHSTETHRGQYFCI
ANNGVGGDLSKAVTVNVN
Ig9
VPVTFHSKYQAQSAILGENTSLVCSAKGETPITLNWTVDQSRSSRHIINETPTRFGRISTLQIMAVEREDSGSYLCAAKN
EFGADETTIELTVKESPE
FNIII-1
PPTNLEVSLRKNQKAFLSWTAPYNGNSPIIRYVVQYKFTSASWNSDVLDATFDAKEASGTIGGLRPATTYNFRVLAENDI
GVSQASEVVTVTTEEEAP
FNIII-2
GGPPQAVNVEALDSQTLKVTWKPPREDLLYGVLRGYQVGHRARGSNDPYAFQILEIPANATPPAELSLNVTELRKYSPYH
IVVAAYNNKGRG
FNIII-3 partial
PLSQEVMVMTAEDNGKPSVKTTSAYKMTVVKLEKFTNYSIQVLAFTKVGDGVKSLPLYCKTHEDV
FNIII-4
PGSPANIRVLPASGESVLVVWRPPLQTNGIVTRYIVYCKNLDGRDVRVEKLVNEPVIRHSVSANVLQHEVRRLRRNRRYE
FWVTAATGAGEGQSSRVITQSPASNKA
Ig10
VPAAVASFDDVVVTNWKENLLLECHTIGAPFPDRRWLIDGHTIVETSRIRIVANGSLSVLDVQGEDEGNYTCRVENDHGH
GEVKYQLIIQAPPSLSSFEVLSITMTSITVHW
FNIII-5
RVKSSGGHPIQGFILNHKRDQETTWERTEISPSQETYTLESLACGTKYHLYLVAVSRVGIGNPAETLTITTEGTIPTIPP
KQKLIEENSSFVTLRLDSWIGNG
FNIII-6
DCPITALSVEYRVKTHYKWAFVTESANLEQKKLVIPGLLPATWYKLRMTANSSAGASTANYDFATLTPSGGTVPPELGLD
EDIKGILFYLDLR
FIIAIAASLFVILAALI
Cytoplasmic tail
TMCICIKKGKASSGKRKEKQINTLHKEETMQSRPNSWTKPPRQDGNADTFSRIHTGNQHIGRYADYEPVIRENGSIPRDG
APPMVPVNKDHLQRTALADEGITRFSGPSFEDPNHHDETAETTFIFENPLDENEHITTGHGTLSGRDKKPQVTCRETSEF
LGPRF

Sm54.2

Ig1
APEPPHSVEFSNSTGANIVCRAEGSPPPSVSWVLADGSGVGNVPNLREVTLDGTLTFPPFRAEDYRQDVHSVDYRCVVTN
PVGVVLSRLVHVKADGKYAILP
Ig2
VMNPVYDLQVYDTYAIKGSSAVLRCSVSSLMTSVVNVTAWIKDSAFKIESSPTPGNGELYIRDVDHNDAQTSYRCQVRHR
FTGETRQSTTAGTLF
Ig3
VTEPQGNVAPKMFESDSKLTALEGESVILPCAAQGFPLPEYTWFLQGRDGQLISIYLGERFTQISSILIIKSVMVSDTGI
YVCRAENNVDIDQTEVSLD
Ig4
VSSPLKATLIPAVQILEIGSSLKLQCKVSGWPVSTIEWVKDGRSLVSSLHLLVANDTVHIDYVRPDHKGMYQCFASNNYD
MVQASTALFMG
Ig5
DRLPQLLEGFQEHTLQIGDSVSIKCSFKGNPIPGVTWRLDNTPLPETHRYMVELQTDGIEKIVSALNVNGVKSEDGGLYT
CEAANKAGKTSHWARLNIYGPAELP
Ig6
QMSVFKTRMFFKLYILISCLGRPAVRSMGKISATSGGNVYLNCAYYGYPIDKILWKKGLKNQNKLISLFSVNLRQDILPN
GTLVISNVQRASDSGRYTCVASNKDGDSASQSLDLAVL
Ig7
VAPNIIPFSFQAEHLYAGVIARISCVVYQGDAPIQLLWLKDGQPFEDSLQVEVKTIDDYSSILTIPEVKPIHSGDYTCVA
KNLAATVNYTAPLT
Ig8
VQAYWMVEPNDTSSFLGGSVHLHCLAGGHPQPLITWLKVQGIVFVWFKFVCCFNFLLFGDDYTIFVNGTLYISHVDDQHN
GYYFCEAQNGIGQGLSKVVRLI
Ig9
VHRPPKFTNPLQRRIIQKGDAIKLECDPKGDRPIQVTWIVNENKVHRKDARYKLRETHSDEKFASELSVDAALRTDSGSY
VCAARNVYGSADAIFEVT
FNIII-1
VQEPPDSPSHIEINDVLNRTLTLMWVAPYDGNNPLKRYVIQYKQAEEQWLDDGKNISVDPTRSKVVITGLQPASDYEFRL
FAINDRGASEASDVVQASMAEEAPSG
FNIII-2
APEEVEVEAADGTTVNVFWKPPSKEHWNGNIRGYYVGYRVAGSGDEYIFHQHDVPEGFTDRLSLVLTELQKFIQYAIVVQ
AFNGEGRG
FNIII-3
PLTEEIIIMTAEDVPSKPPEEVRCSVLNSYSINVTWQPPPDESINGILRGFKILFHPIENGFTASPIIETKSVFTPKITL
KDLKKYTNYSIQVLAYTKKGNG
FNIII-4
VKSEAIMCTTMEDVPNAPADIKALPSSPDSVLLTWKPPLNTNGQITKYIVYAKNVDNESEEKIIHVAKGDITSQEVFGLT
KENQYRFWVTASTKIGEGPKSE
Ig10
VLIASPGDVNIPAKSASFDEITATAVKSTLTLPCKAVGIPMPDRKWTRRGKAVDKSDRLYVSTDGSLTITNIQEKDAGNY
SCKLSNTFGMDQVVYTLI
FNIII-5
VLAPPSKPNLAVTLVTTTTLDVQWKVGIKGASPIIGYLLHFKKEFGTWQVIDIGAREDGHRLYDLQCGTNYHLYVIAVNK
VGQSEQSKTLTLRTKGQ
FNIII-6
VADIPPKEELIEEGSTYITVRLDSWKSATCPILNFVTEYRVRSQNKWYMAPNDIKPDQKRLVIRDLMPATWYVLRMSAFN
NAGSSVVEYTFATLTLTGATVTPIEMINDIPAANVLNLLDLK
IVVPAACTLFIFAIALILICIC
ARRSSKQKHKKKETEKRSENNHSRPPLQKRQKKDDVYVDTGFTYFQRNQPTSSGVQRDNKKRQSRLPREYRE

scf7180001248762

Sm62.1 Transcripts_eggs Locus_3954_Transcript_1-9

Ig1 partial
VDGTPVSNVSGLREALTVGKLAFLPFRAEDYRQDVHAVHYRCVAVNSVGSVISREVQVRA
Ig2
VLLQVYDVHVYDTYVISGNTGVLKCHIPSVLSDFVRVISWTRDEAYII ?SPTHSKGKENDL
GDDWVFGLTDFSCLSI
Ig3
DPKGNVPPKLTDTVTKVKVKQGDILVIPCVAHGNPAPKFWYVKPERNVQHLVHLGDRAYQTASSLVLVDPQVSDGGIYIC
EAQNSVATERAEIKVMV
Ig4
VSLSAKIEPAVLIAEEKQAAIFKCYPSGFPITGIIWLKNGRVLMKINHNISHVSELKVES
ADLESRGMYQCVVKNSQESSQASGELRLK
Ig5
DAAPVLKRGFSDKVLQPGPGFSLQCVAVGSPPPSVTWTLDGITLKSGKDRVSVYTFLDPTGDVVSTFNVSNTRTEDGGLY
RCIVKNKAGIVEYMARVNIF
Ig6
KPAVRKTPKLAVVAGNDIWVDCPMYGYPIDNITWEKGRSLPFDLRQSLFRNGTIKITNVQRGVDSGRYTCIVNNKHGQAA
KEDKQLVVM
Ig7
VPPKIIPFAFVGDQFHLGMRAHLTCAVSEGDLPVRFQWLKDGREMPTTLGVVVRSYDQHTSSFSIEGVSSQHSGNYTCVV
ANSAGTTSHSARLLVQ
Ig7.2
VAPKLVPFSFHGDYLYEGGEARVSCVVSQGDLPISLQWKKDGRRPDLEESVGLSIRVIDDYSSLLTIENVQRKHSGTYSC
IARNEAGVAEYHTHLAVN
Ig7.3
VPPKITPFSFQDAHQGILARVSCVVSHGDLPLKFTWEKDGVRLDSSLGVEVRLFDEYTSVLSIGSVEAKHDGNYTCIASN
DAGSASHFALLRVD
Ig8
VPPWWVIQPQDKSVVLGGSILINCSADGFPKPVISWTKVEVISSSSVLHVFTNGSLWIKQALIEHKGRYFCQATNGIGGG
LSTPINILVH
Ig9
GPPIFDIKYRNQTVRKGESFEVPCEARGDYPINIEWIKDGEERIGSYAVKEGTVAEAPVSHLHVMAADRRDTAVFTCIAE
NAFG
FNIII-1 & 2
EPPDFPQNVSVANVESRDLILEWITPFDGNSRITKYVVQYQPLGGNWQNEKVNEVSVNGKETVAPVAGLSPATSYHFRVL
AENMLGTSALGEEVNVTMAEEAPAGPPESAKVEAVNPQTLKVSW
FNIII-2
APKPEQWNGKLRGYNIGHRIVGNGVDSNYIFRHTDLIELSADDLHTFITDLQKFTQYGVAVQAYNDAGK
FNIII-3
VPSKPPQELRCTTLTSQSIHVTWQPPPSDSTNGILRGFKIFYKPIKEWDDATIQEKIDTPKKTLKNLEKFTNYSLQVLAY
TKMGDGVASSPIYCRTLDDG
FNIII
VPPAAPALQIVATTPTTIEIAWSLPDDGGIRIQGYFLFIQARYIFSGYTLNLKREFGQWEQMTLSPEVRNTTLTNLECGT
RYHLYLFAFNKVGMGSPSEVAVPSTEGT
FNIII-6
VPNVPPMETVIEPGVTSVTLNLYAWKNKECPIQYFVVEYKLQMSSDWSFVSNNIKPEQRKLVIPGLIPAMRYDLRMTAHN
TAGSSVAAYKFATLKVDGG
IVIPICISVTAICVILVVSCLCM
Cytoplasmic tail
RRRKGTSRRRVKAIPRFLPRAYGSKKHSSGHGHSGHQHKKKPAIQKPPRKGPVKKKAKVQSAGTISREFGVP
PPIGRAVQLDKQFSSNSSNGGTFVRRRLSDAVCPIGRLSQNFRTNDQFDKRRMTEVSEAFADFVYVYVSDAV
CPTERLSEIVRCHLYETGSTDGDHPEGPTFIKEPPNNVDF

Sm62.2 Transcripts_eggs_Locus_4384_Transcript_2/2

Ig1
EPPNNVDFSNTTGVTIECLATGDPLPHITWESSTGTSIGSLDGVREVLPNGDLVFPPFKAEEFRQDAHAAVYRCAATNAA
GTVLSRDVHVRA
Ig2
VVPQPYDIQVYDVYAIVKTTAVLKCHVPSFLEEHVRVTMWLRDEVLIIEPTSIMYAVLPSGELHVRDVEMEDSGSSYRCH
VKHSLTGETHFSSASGRLYV
Ig3
EPQGSVPPKITDIMTAVHVTEGETVILPCVAQGHSAPKAEWFAKLNKRQLFPLHIGERVQQTSGALIIHRAQTSDSGTYV
CDVSNEAGNDRGETTLTV
Ig4
ASLTVSILPEKQVVDIGKSATFRCVVTGFPVVYVSWYKDGRKLQSDAEKTLSDDTITIASVHVTDKGMYQCLVSNNQESA
QGAGQLILG
Ig5
DAAPELVGVFDDNTLESGTDVSLACVATGSPAPVITWFVDDVSLQTSERIRVVGRADTDGSIVSVLNVTETRWEDGGTYR
CLANNKAGTVEHIARLNIYG
Ig6
RPAIRPLDKVTAVAGEDVRLNCFYYGYPVTSINWEKDGRALPFNLRQIAFPNGTVVIRKVQRATDSGKYTCTVVNDKGQS
AQEHVEMAV
Ig7.1
VPPKITPFSFQDELVREGMRARLQCVASEGDAPLYLRWTKDGNPLQESGLAGVTVRDLDDYSSILSISHVTPRHNGNYTC
IATNEAATAQYTAQLSVN
Ig7.2
VPPKIIPFAFLDDQFYMGMRAHITCAVSQGDLPISFRWLKDGQELPSALGILTRNYDEHANSLHIESVTSKHTGNYTCIA
ANMAASINYTAQLLVH
Ig7.3
VAPKIVPFSFPSDHLFEGVLARISCVVYQGDLPLSISWHKDGVPIPVSDHVVVREIDDYSSILTIESVLQSHSGNYTCLA
HNSAATVNYTAELVVN
Ig7.4
VPPKIVPFTFQDEHLLDGMLVRVSCVVSRGDLPLVIRWEKDGLPIVPDVGGMSVRAFDEYSSVLSIDPLSRAHSGNYTCI
ASNHAATATFTVPLVV
Ig8
VPPRWTLEPRDSSMLLGHSLQLDCQANGFPEPKVTWMKTQGATVGEFVEPVEMSPDLSILSNGSLLFIRAREYHAGHYFC
KASNGIGDGLSTAVHVTVQ
Ig9
VPPRFDVKFVNQSLKRGEGFRLECPPNGDKPMSFMWKKDGVLLDPLADTRRYKIESPADRGVSDLTVEEATRIDTGIYNC
KAENEFGTDDTNLQIT
FNIII-1
EPPDAPGDIRIQNIGSRNVHVTWSHPFNGHTRIIKYVVEYNQGPEEWEDGLVELAVSGWDTNVTLHGLRPATHYQLRMFA
ENELGLSSPGDFVTFLTKEE
FNIII-2
APAGAPVDVEADAVDANTVVVRWPPERHVWHGKLRGYTITYYKKLDSTDAPNFRTVQVEDGSEGNHSAFITDLEKFTKYS
LTISAFNDQGQ
FNIII-3
EPPTDVQCLAYTSQSIYITWQAPLSTSFNGILRGYKVFFAKADDTAEDGSLEFKSTTVVRTTLHGLEKYTNYSLRVAAFT
KVGDGAA
FNIII-4
VPDVPADVKAIPASKESVLVAWKPPLHSNGVVTKYYVYAKHENDPEEDAMKHTAPPSALYHEISGLKTDQSYEFWVTAAT
MIGES
Ig10
VVTTLWKKYIKLPCKAVGNPATERMWTLTPVQSTRLVWEFVWTNANDEDFVYGHCRAHTVKESDRLRILPDGNLMVKNVQ
WNDSGNYTCQVKNGFGYDQIV
FNIII-5
APPSAPIINVLETTPTSIVLEWKLDTDGGTNVQGYTLNYKRESGHWEQRDLAPDSDSYTLSGLKCGSHYQMYMTAFNKIN
TG
FNIII-6
ADVPAKEDLIDEDVDYVTLDLHVWHSDSCPVSTFKVEYKQKNSLNWHLLTDELKAEQEKYVIRNLTPGTAYLLRVTALNG
AGPSVATYDFITLSRQ

Cytoplasmic tail
AIPRFLPRAYGSKKHSSGHGHSGHQHKKKPAIQKPPRKGPVK
HEDPRGILFYLDLRIIIPLAALVIVVILVLSVTCICANRRGNAAR
GMDSQNGPHRPCLRTDSTLSSFAYLQRHRSSCDGSILRSDTQEKSIPYSTYNLPNLRSSAEFKTFGQRNSVSDPPPLPPQ

DEADAQALQHFSRPPVTNLQSPNGKMQRTGYPVAIPTAPTPPPQIDANSADQTGSPGLKKVPEVPPKPNMVTMETRFSSS
PDKVRVRVLPGA

scf7180001237055

Sm55 Transcript_Locus_8916_Transcript_1/3

Ig1
EPDDIIEFSNSSGVQIDCNAEGTPQPTLGWQLADGSSISNITNLRLVYPNGTLLLPPFKAEEYRQDVHTATYKCTASNLV
GTIFSRDVAIRA
Ig2
VTIQPFDVYVYNVYIIRGNTAVIRCHVPSFLTEYVEVLSWIRDGAFTIQVNGIEEGKYAVLPSGELLLKNAGPEDAQTTF
RCQARHKLTKEIKTSVSAGKL
Ig3
EPEGNVPPRMLIHQTSVVAEEGHDAIIPCVSNGHPVPEESLSSVIKPINEDARYKIVQGALIIIDVQQYDAGKFICESTN
GAGTQQMETELIVF
Ig4
LFATITPPIAVIDIGHKTTFNCEVTGYPIKNIQWLKDGKQVSHFENQLNKN-NATLTVKSVSAADKGMYQCF
VKNDFDVVHASARLKLG
Ig5
DSPPEFLTTFEEKTLRPKESLSILCEATGTPTPDISWTLDGMAVRGNSQKESADSIISWL
NITSLKVENGGVYTCHAKNRKGSVEYSTRIYV
Ig6
KLEARSRPKISAISGDTVWLNCPIYGYPFDTLTWEKDDILPAHLRQVILKNGTLRLENVQRGRDEGRYICSVRNNNGESA
RGYVDINIL
Ig7
VPPKITPFFFQEDVVRQGSRARLQCVVSDGDTPMTIKWLKDGSEISKALGITIREIDEYSSILMIPAINPQHNGNYTCVA
INKAANTHFTTKLAVN
Ig8
VPPHWVIKPQDMHALVGSGVIIDCMAEGFPKPTIQWMKTRSLYLRQKTSSNFSNKILVATNGHIQILQNGSLRIPYLSEH
NEGYYFCHASNGIGDGLSKAMYLKIY
Ig9
PTRFIFKNNNKSFSISEKIHLICEATGDLPITFNWKLNNKTLDIHRNLRQLRSKENATKNSAISELFISKAKKNDSNTYV
CIAKNDYGSDETNFHVTI
FNIII-1
VPDPPVLIEVKSSGNGSNLVKWQIPQDGSSPITHFLLQYKEKKDWNLSQNLSLLRTRNWTLINDLKPYAFYDIRLAAANS
IGYSNFSNDLDFQTEEQ
FNIII-2
APSGPPLNVEIEPVDKQSLRITWKQPEKKFWNGVIRGYRIGYKVSRSDSTYTFISIEIPEDYTDDLIYQLTELDMYTQYM
IIVQAYNGKGNGPPTEELH
FNIII-3
VPSVAPYNIRCSPLSTTTVYIIWDPISPDYTNGILRGYKVFYKPFDDWYDAAYHSKSIDVPKLTLQGLEVNTNYSIEVAA
FTKVGEG
FNIII-4
VPSAPGDIKVLSSSSDAVLVTWKPPLKPNGAITKYNVYVRAIDYDEEIGAETRYSKDAVPKPGGVDDTDIHTSRDDLVYQ
ITGLDKNQRYEFWITAVSGVGEG
Ig10
VPAQTAAFDDVIKTPWKKDLKLNCRAVGSPQPTKSWIKGGKTITPNERAQLGPDGTLVIQRVDLEDSGNYTCKVQNKYG
NDEVKYLVVVL
FNIII-5
VPPSAPNLGVYSKTMSSLQLKWHQGSNGGDPVRYMLYYRKDGDKWEKKELFADQDSFMLENLACGAVYNLYMESINNIGV
GESSDVVTITTDGDA
FNIII-6
LEPPTKEDLIHEGNTYFSVHLDAWPSSGCPIKNFTIEYRKADTVTWTMVKTGVKPIEKRIVVSGLTPATFYHVRVTAFSN
GGPTVADYITATRTSTGG
TMSPSELEHQEGDRVIWLSLDWILITASIVLAMILVVVVVTLICIKRWTTNNQKSA
Cytoplasmic tail
RKSHHMSVESFVHLQGNNPTNSRVYASLKKGTLRKDNLSPYASSTLPGCSPDIRSQGRLSVPTGAHSG

Sm41 Transcript_Locus_5994_Transcript

Ig1
VDPMGPVFIYQPPNSVDFSNSTGASVECSAHGNPLPVVQWIHAATSLPVTNVSQLRLVLPNATLVFPPFGAD
HYQAEVHSSVYRCRAANLHGTIISRA
Ig2
VVLQAYDAQVYDEYVIRENTAVLKCQIPSFVADYVTVTSWVRNSMDNIETDVKKGKFFVLPSGELYIHNVSA
QDALHTYHCRTLHWLSGEAKLSATAGKLVVT
Ig3
DPNGSVPPRITDGKSIVQAKEREMVVLACAAQGHPAPSYRSVVVRYSWHHKVDSGQQVTAISESRLHQVHGL
LIIDHVQPEDAGTFVCSASNSLGTERIETSLIVN
Ig4
VPLSVHIEPVQQILDRGRMATFTCVISGHPISSVIWLKDGRDVKKENILRRDMLQIERVHREDSGMYQCFVT
NNVETVQSTAELRL
Ig5
DSLPEILSAFKSHTLNPGVSLSLRCVAAGIPVPKVIWNVDDTLVNPGDRIRVGEYNDMNGNVISHVNISRVQ
VEDGGLFACTASNKAGNTTHTAPIAVYG
Ig6
PYIRSMPKITVVAGEDLRMRCPVSGHPIDSIYWEIDGSRLPVNHRQKSFHNGTILVQSVQRNLDAGKYTCVA
SNNQGNTARRDFEVAVL
Ig7
VPPKIIPFSFQEDQTYEGVRASVFCSSSQGDLPLNIKWYKDNTLVQPKSDVTTQTIDNYASTLVIELVKAHH
SGNYTCSASNAAATVNHTALLIVK
Ig8
VPPRWHVEPVDTSTALGSAVQIECQAEGHPPPLISWFKSLVADVSNFTLIFFFLLLPDVSSSDFVELTATSL
SHQGSLRISSALEYDEGHYMCKATNNVGAGLSKVVFLNVH
Ig9
VPAKIDVKLKNESVKMGEEAKLRCNVHGDLPIQVTWSSSKHAISTDKRFAEDLKSDVGMTSMLKIMNVVRKD
STMYKCNAKNQFGEDEASVLLIVQ
FNIII-1
ELPEAPNNIHLVEEGSRAVHISWSRPFDGNIPVTGYLVQFTSGSDWSSHVFNLTIPSTQMRVTIKDLKPATK
YRFRVFATNELGMSESSVLVSTTTGEE
FNIII-2
APSGPPKDVRVEAMNSQTLRITWKPPKQEHWNGEILGYHVGYKLYNYSEPYNFRTGSPLNLMLTFEKLKKFT
KYSIVVQAFNDHGIGPNSEEVVAMTIE
FNIII-3
VPSSSPGNVKCSAVSSQSLHIQWDPPPTHDINGLLQGYKVLYKPMREWASRGGSVFETKITPALKTMLHGLE
KHTNYSVHLLAFTHVGDGVKSEPVFCSTLE
FNIII-4
VPGPPADIKALTMSLDAILVSWLPPVKPNGNILKYTVYVRLTDSGREETTKVSVPDSMLRYESKGLSKNRRY
EFWVTASTAIGEGESTRVTTQTTSGPL
Ig10
ARIASFGNHVIVSWKQKLELPCDFIGTPAATVRWLFLGETLQTSNKLHVLSEGKLIIKGIQSNDAGNYTCTV
QNSLNSDNITHVVVVE
FNIII-5
VPPEAPIVSIVSTTMKSIHLSWTRPLDEMRSNDIYILSFRQDYGQWSELNLQPSVYTHTLQSLTCGTRYQVY
VTPVNRIGRGQASDIITAKTKGD
FNIII-6
APKTPSKDKLIITNSTFIILNLDSWFDGGCPILYFVVEYKRKETSDWTLVSNNVKPDNKRFVITDLAPAEAY
QLRVTAHSSAGSNFAQYDFMTLIEETD
DNRRHREISMQEEEGSLSFLDLSIMIPAVASIFLVAALISIVCICL
Cytoplasmic tail
KRRKMESGNTADKINPKYQSVNPDTKKKTYTETVPTIENGSLQDIIPAYPSNATELNYNSKYPPAENLYNEN
LKQAVHMCPRHQISHTEAEVPDLLEHGPDSSSSEDASPQFHHRTRRVDSYPRHPGDPHPLGYPHHTLNRNR

scf7180001248653

Sm53.1 Transcript_eggs_ Locus_449

Ig1
SVELHGPVFIQEPTNHVDFSNTTGARVECTAHGTPLPAVQWLLADGSPASDVPTVRVVYSNGTLAFLPFPAE
GYRQDVHAAIYRCRASNSVGAILSRDVRIR
Ig2
VVMQMYEVQVYNEFVIRGNTAVLKCHIPSFVTDYVKVMSWVRDTTFNVLSEVETGGRYTIMPTGELHIREVG
PSDAYHSFRCRTIHRLTGEIRISSMGGKLVIS
Ig3
DPQGSVTPRITDSKTLVQIHKGEAALLPCAAQGYPAPTYWYVKSNRSQMNSLVLTDRIKQIGGSLLIQNARI
ADGKTYVCVVSSNVGNQSAVTVLSVT
Ig4
VPLSAYIQPHKLKVDVGSSAVLNCKTSGYPVASLVWLKDAQLVRPQPHPPHSLHIETLRREQRGMYQCMASN
DHEAAQGTAELRL
Ig5
DAASDIVEGFQERVLSPGTSTSLRCVASGNPAPQMMWMLDDTPLVNSLHTELVSAQGEVVSYLNLTGVRTQD
GGDYTCLASNRLG
Ig6
PPGIRPMSRMSVVAGFDVTLKCRVYGHPLESLSWEKGLLLPVNRRQKLFPNGTLAIQNIQKTIDGGKYACVV
RGVTGEAVRKDMEITVM
Ig7
VPPKISPFSFQDEDLYEGMRAQVTCAVRQGDLPMAIHWMKDGVPIEATSLGRDGALVARTFDVYTSSLSIDS
VASEHNGNYTCVASNMAAAVTYSSSLRVN
Ig7 ?
APPVLASFSFPSVGIHEGMVARVTCSVTQGDLPIYFHWAKDGRQIASGEGIAIKDFDEYASILTINDVRHRH
TGRYTCIANNAAASIKHSTHLIV
Ig8
VPPRWLVEPKDTQVLVGASARMDCQADGYPEPSITWTKAVGKHPFSNLPIGVCTIRINKPHLAINKLTKRKK
ALLFITASFRLDFDDFANRKNLLN
Ig9
PVQFEVRSRNQTAKRGENVRLQCNAKGDSPIKVTWSVNSHPIEPAARPRYKIKEMSSKHGFLTELIVTRSER
SDSGVYSCSATNPHGRDSTVIHLTIQ
FNIII-1
EPPEAPRSVNVEDYDARSVNLAWLQPYDGNSQVAKYIVQYKHADWIGGSANETVIGRVTSAVVSSLLPATKY
AFRVMAENEVGVS
FNIII-2
APSAPPTHILIEATQPQCLKVSWKAPQKDLWHGEILGYNVGYKMQDSSKPFLFKAVESASLDGGHLELRGLL
PFTKYDIVVQAY
NRVGPGPLSDAIGASSAE
FNIII-3
VPSRAPDDVRCSAHSSQSVHVTWAPPSPQSVNGILQGYKLLFREIHEQKRHQTLGHIMETKITPSLETILHG
LAKFQNYSIQVLAFTRVGDGVKSDVITCQTFE
FNIII-4
ALVASEDSILLTWLPPSQPNGIIIRYTVYIRTIDKDKETTKMIVSGSQLSYDFKGLVKNHRYEFWVTSSTSI
GEGQSTKMVSVTLTSK
Ig10
VAAKIVSFGASIVIAWKEDVHVTCDAVGIPPPTRVWNVGYAKLIDRGQPLPQLERYQVQPDGSLLVRNVQLT
DSGNYSCRVENGHGSDINFYIILVQ
FNIII-6
PSAPAKEQLIYPNAEFVALNLNTWDDGGCAILYFIIEYKAKSSPDWVLVSNNVKPQRDAF
LIPDLESRVSYNVRITAHNSAGSQTEV DFATRGR
SDIAQDDKSDMDEEDQAPFYADLKLIVPIASAILGLLVIFATYALCI
SFRRNKLHDKNESENYSQPEKDVGSSLAHEFDKNSVQTPVSTRCNYLQQEFDHFPGSLAAPTRPKLNYTQTSLEDVCPYA
TYRIPESSNKAQVHTSAWHQLPLKTQDFLSGGREKRDCGKQTSSNKNFSPQVTDQESTNYQNPITSCDQINHQTSSNYPI
TSPNHRQVAALNLEANINGISLFRLWATFIAVNNTIPFYRPWIDVFVCTSDIVA

Sm53.3

Ig1
FVQEPPNRVDFSNTTGARVDCTFHGTPTPAVQWLLADGTPALDVPEVRIVFSNGTLLFYPFPAEGYRQDVHAAVYRCRAS
NSVGAILSRDVRIRA
Ig2
VVLQIYEVQVYNEFVIRGNTAVLKCHIPSFVTDYVKVVSWVRDTTFNVLSEVETGGRYSIMSSGELHIRDVTPNDAYHSF
RCRTVHRLTGETKISSMGGRL

Ig3
DPQGSVSPRITDSKSFVQVHQGDSAVLPCAAQGHPPPSYSWFVKSNRNHMTPVVLSERIQQIGGTLLIRNARIADSETYV
CVVSSNVGNQSAVSVLSVT
Ig4
VPLTVYVQPHKLKVDVGGSVTLTCKASGYPIASLVWLKDAQLLRPQPHPPTALHIETMQREQRGMYQCLATNDHETAQGT
AELRLG
Ig5
DAAPDVIEGFKEHVLSPGSFLSLRCIATGNPPPKMMWVLDDKSLEDSRHARISQRVGSQGDTVSYLNLTGIRTEDGGEYS
CVAVNRLGNATHAARINVF
Ig6
GSLGIRRMTPMSVIAGEDVFAKCRVYGHPLESITWEKDGLLLPINRRQKIYPNDTLLILNVQTSDSGKYMCVVRGSGGET
VRSTLEITVM
Ig7
VPPKVSSFSFQDEDLYEGMRAQVTCAVRQGDLPLTIKWLKDGVPIEDTPPGQRGTLVARVFDGFTSSLSIESVASEHNGN
YTCVASNMAAIVSYSTTLRVN
Ig8
VPPRWMLEPRDSQVLVGGSARMDCEADGYPEPAITWMKAVGVPVDFREIVANGVNVMVFANSSLLLTGVKESDQGYYLCQ
AANNIGEGLSKIFFIDVH
Ig9
VPVSFDVRSQNQSAKRGENVMLKCNAKGDLPIKLEWNVNSHLIDPDFRARYKIKESNSAHGLVSELIVTRSERADSGFYV
CLATNAHGRDDTTTIHLAV
FNIII-1
EPPEAPRSLNVDDFDARSVHLIWSHAYDGNSPVLKYIVQYKHVLAEWFGGSANETVKGSTSGAVVSSLLPATKYSFRVMA
ENDVGVSEPSETVVVTTAEE
FNIII-2
APSAAPIHIHIEATQPQCLIVSWK
PPQKDLWHGEILGYNVGYRVQDTGEPFLFKTVEIETDGPGRLELRGLSPFTKYDVVVQAY
NKIGSGPISDPIAAATAEE
FNIII-3
VPSRPPSDVRCSAHSSQSIHVTWSAPTTSSIHGVLQGYKVLYKSNSQEQHRVPFHGQHFPSDLETKITSSLETILHGLTK
FENYSIQVLAFTRVGD
FNIII-4
VPEAPAKVKAVVTSEDSIFLTWLPPYQPNGLIIRYSVYIRTTNDDVEVNNTTKTIVAGDQLRFDIKGLSKKQPYEFWVTS
STSIGEGQSTKMVSVIPNSK
Ig10
VAAKISSFGTTVVTPWKEEVVMECDAVGIPPPTRVWNVGQALPQHDRYEIRPEGSLSIRNVHLTDSGNYSCRVVENTHGS
DEIHYAFIVQ
FNIII-5
PPVPPHLVVSSTTSNSVTVYWKPDANGGSPITFALTLKREYGEWEETQLEADCRSHVIDNLWCGSRYQLYISAANSIGAG
EPSEIASFKTKGS
FNIII-6
PSSPSKEQFIYPDSEYVALNLNTWNDGGCSILYFIVEYKPKTVADWILVSNNVKPQRDTFLIPDLEPRVSYNLRVTAHNS
AGSQTQV

scf7180001248602

Sm91

VFTTGPPLRVDFANSTGARIDCTARGNPSPLVKWTLLDGNSADDIPGLRHVLSNGSLIFL
PFRPEDYRADVHSVTYICHAKNTWGTIRSRDMQVRA
VVLQLYEVQVYDEYAIRGNTAVMRCHIPSFVKDYVSVMFWLEEPASGGSTNVIET
GGRYLITATGELHIGHANTSDNANAYRCRTLHRLTGEMRLSAVSPSGRLYVT
EPRGSVPPRITDHRPSVHVVQGDAVLLPCAAQGFPLPSY

VFRRELPTTDLRFTWSKEMRFCCRAQLKASHYHLIGEIYESLATCGQLGLVCCVGLCRRL
KIRIILSWFAKMNEVEVLPISTSSRILLMGGSLMLTSALILDAGTYICEVSNSMGVVRIETILTV
APLSAYIYPQRQVVDVGQSATLTCVISGYPFTQVTWMKDGRPLLTDTINQLSQEVLRLDA
VHRRDRGMYQCFVNNHLEVVQGTAELILG
DILPEFQRVFSDKVVQTGSFVSLECAVSGSPTPQVVWTLDGQVLKNRNNKVTSANFVDDN
SDVMSLVNISQVGVADGGEYVCTASNRAGSVRHVGRINVQG
GPPLIRLVSNVAAVAGTDLVVRCYVSGFPIDSVHWE
DDRMLPFTIRQNVYPNGTLLIQNVQKALDEGQYTCAAKAGRLVDRKQTNISV
APPKIIPFSFQDEHLREGTRARIQCVLSEGDLPIAISWLKDSRSISAQLGILIRDLDDFS
SMLTVNNVSSLLHNGNYTCVATNTAATANYTSELSVN
VPPKILPFSFRDVQLQEGMRAQITCAISEGDQPVRMTWLKDGHPLNSALGVVVREFDEHT
SSMSIERVFSVHGGNYSCKAGNRAAEVQHTAQLLVN
VPPRWLTQPQNTEVILGGSTYLSCQVDGFPKPTVTWMKAVGDAPGDYRDIAFELLHFKLN
EEGDLQVLGAEEADKGYYLCKASNGIGAGLSEVVYLSVH
VPAYFQTKTRNVTAKMGGRAELVCEAYGDKPLTISWSAHRDPARADALS
YNVNDNYWEKGTISELVIEKVEKSDSGVYPCVATNAYGEDESHVQLIVQDVSDAPLHLRA
SDIGSRKIRLAWTAPFSGYSPINLYILEIKDKSEDEWKDGRNLTVSGHATECIVE
ALEPAKSYHMRLYAKNEIGTSKASKHIEVTTNIE
APGGPPLEVRVEAVDSTCLNVYWKPPRSDLWHGKLTGYKIGFRQHEIKEIQFRVVRLEDN
ENEADEFVMRLTHLKKFTKYRVVVAAVNQMGDGPFSDDILARTAE
PSRSPEGLQCSPISSQGLSVSWDPPPTNSVHGQLQGYKVLYKPVSEWY
DDMPTEVKISQTWKTTIHGLEKYKNYSIYVLAFTRVGDGVRSEPVFCLTKED
VPDAPAAIKTLIISSSAVLVVWKSPLRTNGIITKYVVFMRNSDSGND
EIRKFVVSSNKTLMYEIGNLKKNHQYEFWVTASTSVGEGASTKAITQIPSSR
VPAKIAAFDETVISQWQESITLDCYSVGNPTPLIEWRL
NVQIQVTKRFEILPTGSLFISQLQNSDAGLYTCRVQNIYASDAVVYTLKVQ
GPPQPPRITYLKSTFSSIHVQWEVSTDIGNPVEGYIVYYKRDFGEWESVQLGSVEESHSL
DDLWCGTRYQLYIVAWNKVGIGEANEIKSIRTQGS
APELPAKHKLVHENVSSIGLNLSSWENGGCPILYFVVEYQPVNHHEWMLVSNNVKVQQFL
ILDLAPATKYVLRVTAHNSAGSTIGVYEFVTKPHG
VDILDEVTNEMSPNSGFYLDVNITFPIAALLSFLIVASTCVICCRRYRVNNSIEGS
DGGSEPKRYEARYVVTGDKKSCILGNEIDKGSFTCLLVNTEGADTSSGNTPRNAKR
VAPRGEIQPYATYQLPECCTDAFTPDDWKRFEIYNPGHVPMRA

Figure S3 Maximum likelihood topology of the nucleotide sequences of the duplicated exons coding for Ig7 in the different *S. maritima* Dscam homologs. Support values at nodes are bootstrap values expressed in percentage relative to 1000 replicates. The tree is rooted for convenience at the midpoint. Each exon duplication was numbered according to its physical position in the locus.

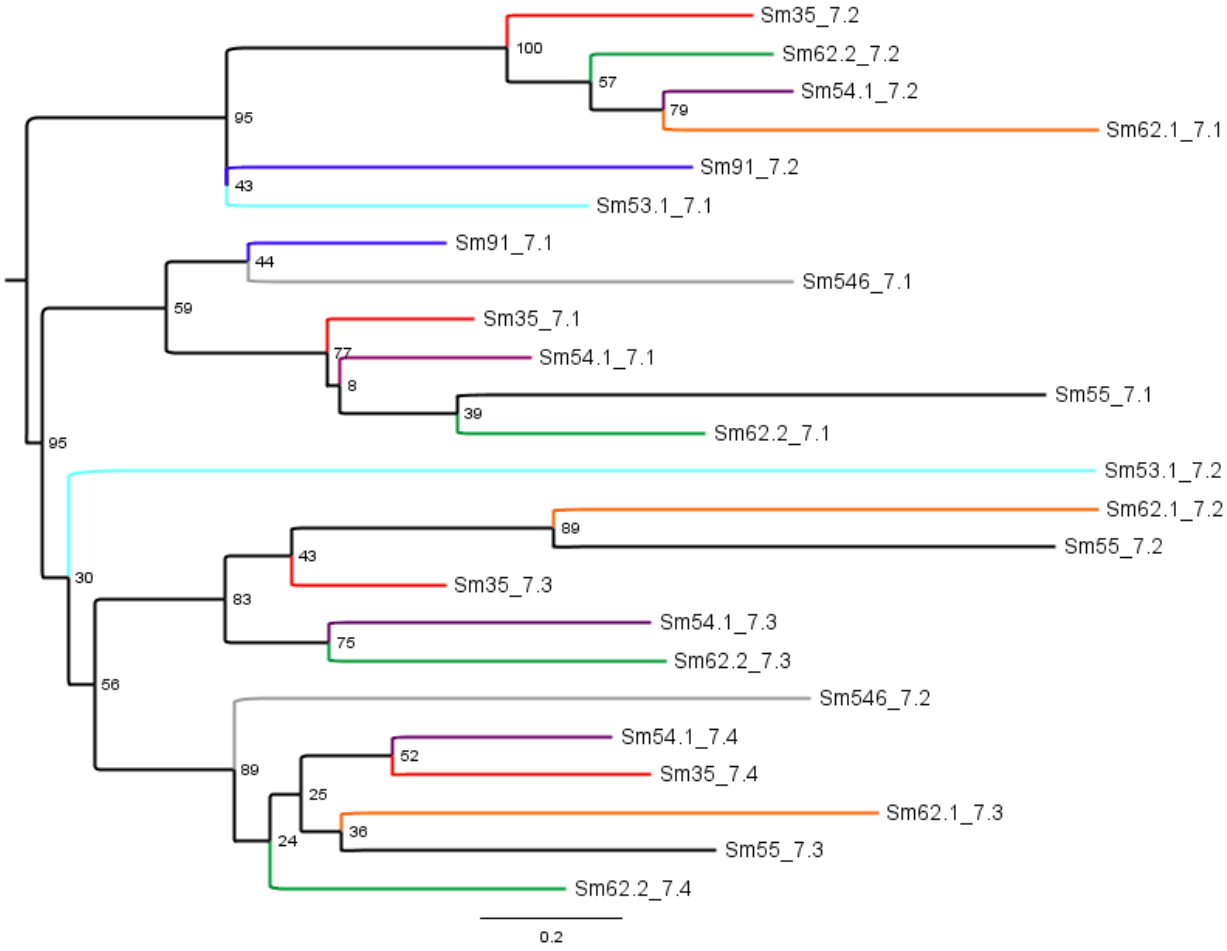


Figure S4 Number of amino acid substitutions per site calculated with a pair-wise analysis of the poisson corrected distance among different Dscam domains of paralogues. A) *S. maritima* paralogous genes containing exon duplications coding for Ig7 (n=8). B) *S. maritima* paralogous genes containing not contatining exon duplications coding for Ig7 (n=5). The comparisons of the different Ig7 coding exon were made based on the groups obtained in Figure S3). Genes *Sm53.3* and *Sm91* were not included. The bars indicate standard errors obtained by 1000 bootstrap replicates.

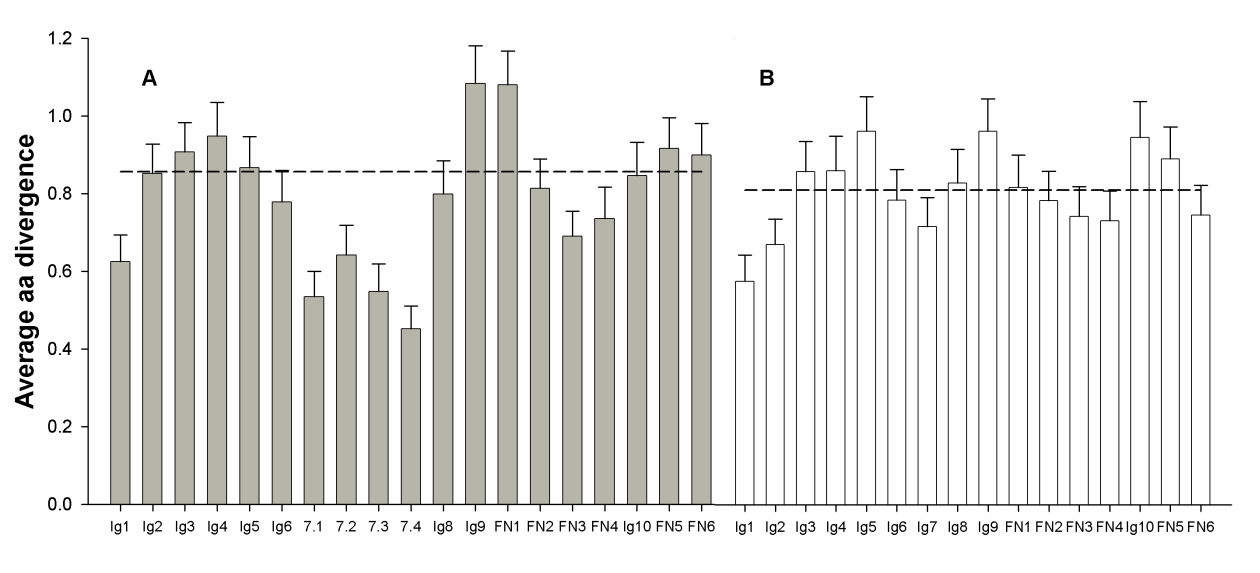
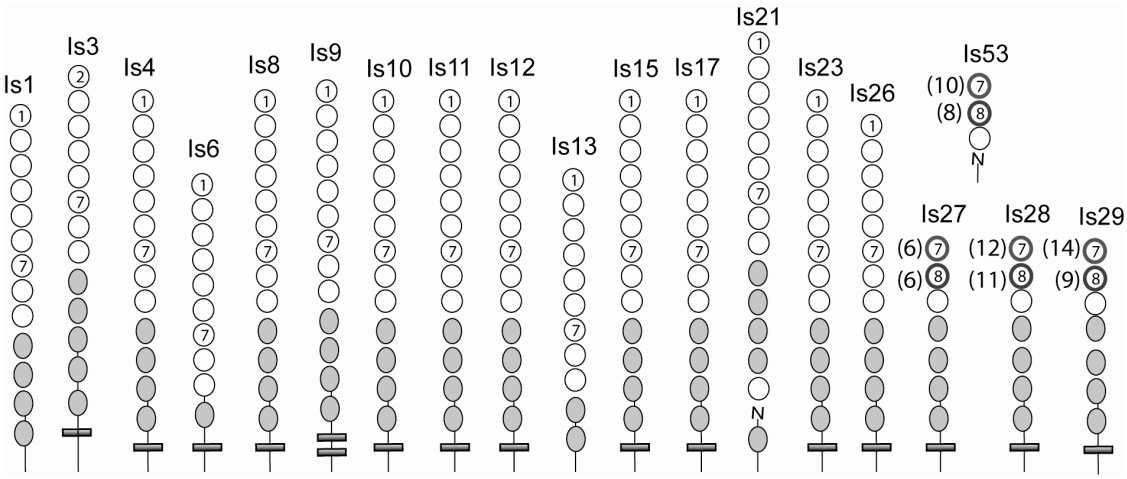


Figure S5 *Ixodes scapularis* reconstructions of Dscam homologues. The round circles represent Ig domains whereas the grey ellipses represent FNIII domains. The Ig7 and Ig8 domains which are coded by several possible exon are represented in bold. The number of possible exons coding for those domains is indicated in brackets.



Is1 contig 979349

Ig1
GASLPGDAPYFVREPPARLRFVNS**T**GAALFCAAKGHPVPDIHWVTVESEVRAEPRPALDI**PG**L**R**RLQPDGTIVFEPFRAE
QYRQDVHSAVYRCT**AA****N**RV**G**VLG**S**RDVQVRA
Ig2
ME*VVQHPYKPQVFDEFVIS**GN**TAVFR**C**SV**PS**F**V**KDFVDFVS**W**HRDDGLTITSTSD*RVRISDFIPKCVFLAGKYSVLPS**G**E
LYVRNTGPS**D**RLRSYHCKTKHKLTN**E**VIV**S**ASSGRLFLQ
Ig3
APQGVVAPKITDSHPFVQLVEGQDIVEIACAVQGFPVPSYSWYREVDGRLVDLSLDPRTAQVDGSLFLSTPVVRDAGKYF
CVVNNSVGEEQIRTTLSVT
Ig4
AALKAELQPTVQTADV**G**HPVTFNCSASGQPVRSVSWYKDQQRLQPTGRISLLASGLVLRIDSVLRQ**D**A**G**V**YQC**YLHNEAD
SAQASAELRLG
Ig5
DVAPFLASTFAEQTS**L**PGNSASLRCSAS**G**S**P**L**P**QVT**W**T**LD**GGGVPDHPRFRVGDFVTSDSTVVSFVNVTELRVEDGGEYV
CRATNVVGEAKHAARVNVH
Ig6
GPPMIRSMGNVTVIA**G**RSFQTVCPVAGFPIHSVV**W**LKGDAKLPTNHRQQVFHSTLTVHNVQRASDEGEYSCVARSGNLSA
RGNTFVHVQ
Ig7
VPPVIDSQMLPDMLTSNQ**G**MNVKMLCSVVQGDPPISLR**W**MHGAQLVSRSSSVSLQSLDDSSVLTIKGVSMRDSGNYTCEA
SNAALTVNRTVTLVVN
Ig8
VPPMWTTEPTNGNVVVGETVVLDCAADGFPVPRIAWKRAEGNEPRNFERLTTSYRVQMLSNGSLVVQDAEISDSGFYLCE
AH**N**GI**G**AGL**S**RVVSLSVN
Ig9
VPPSFSTKFSSQNVKRGQDAVLRCDASGDPELIIMWEKDKQPIDLTIEKRYSLFEETLDGRLGSSLTISSTERWDGALYT
CIVRNPFGSDETNVQLLVQ
FNIII-1
EPPSAPTEVKAAKIAS**R**TVEIM**W**SPSYN**GNSP**IRK**Y**HVHFANRTTSWDSSSSALLLSVSGTENRATIYKLYPMTTYRIRI
TAENHLGHSPPSDTLEVTTTEE
FNIII-2
APGGPPLHVKV**EA**TGSQSLK**V**T**W**E**PP**RKELHYGQLR**GY**YIGYKEEGKMEAEFQYKNVEALDLNTVSQRQLMSHLTN**L**KRK
TS**Y**SVKV**QA**Y**N**SE**G**A**GP**M**S**DEVSSTTLDA*GMHRSLNCLS*
FNIII-?
GIGPPSEVLTVITDGEVPAT
FNIII-2?
PPQSVKVSAVGSKKVEVAWKPPPLHLQYGDIQGYYVGYRVHGTTEPYVFKTVTRASGPTTQCIIDNLQRATAYAVVVQAY
NEKGAG
FNIII-3?
PLSDEIMVQTHEHDPPPSPIVTVVTRTPDAIELAWTPQEENKDAIEGYVARFRLHEGMDWNEVSLGPDKRGYLFEGLVCG
SAYYFSILS
YNRNGRSEPGELLQVKTEGTVPQPPSHRTGIVPNVTGLSLALAPGGTEAAPSALLHQYSHATTPSGHYLSSRVLPDRDSV
AI

Is3 contig 922315

Ig2
NYVVTSDGQLYIREARDADELRRYRCHTENTLTRRKKTSVNFVRLLLR
Ig3
GELTSPAPPVFRRLGVAVRRDVAAFFCPRRCSPSARSWFKRQGQRMAPVEPSLGRRQVAGVLQFRSAREEDAGTYVCVVA
NSVGEAQVDLELVVTP
Ig4
QAWVAVSPAHVRAEVGHAAAFRCNASGPGLEEGSVEWRLNGRRQATRSRVLHVASIRRQDQGMYQCFVRITPQRTVHAAA
ELIVGD
Ig5
QAPKLRSTFEETTVRPGKPVTLRCVATGDPPPRVTWTLDSTWPIDSRHGRLRVRTSGPDSGTTTGQTGGGSEVVSTLSIS
SAEVQDGGSYACEASNYAGVARHVARLN
Ig6
VYGSVFVRPLNNVSALAGSVFAVQCPFGGYPFGHVYWEKDGRRLPLNQRQTAFPNGTLVIQGTDREPDQGQYSCTVHGPD
SQVVHRSISLHVRS
Ig7

GPQITPFSFQSNLHEGVRAGLTCLVHAGDPPIKIEWLRDGKPLAPGAHPSHDVSVLSPEGGFVSTLTLQRLSSKQNGNYT
CRATNQFASAEYSAELLVKV
Ig8
PPSWTLEPNDTTAVSGRSVFVDCQASGVPQPHIRWKSAAGTGSSFFGFASDLRPDIKLNFFFVQRAPASEEYTFVAPSTP
AHNQDLLGHCFCSLHR
Ig9
APRFTAKFTTVAVKRGETAEVSCPAQGDLPIRFHWLKNNLPLNILKEHRYSRMEDRSDDVTVSKITIQSAERSDNAVFSC
QAANEFGEDSTNVQLTVQ
FNIII-1
DVPDA**P**ADVDVREVGSRTARVT**W**SAP**F**N**GNS**AITQ**Y**VLHWKTADGL**W**QDTMSVSGMETK**A**TIRSLL**P**TTA**Y**QLRV**R**ADNV
FGN**S**DFSVVTEFTTSQE
FNIII-2
P**P**RFP**P**K**N**VQATASNSRAIS**V**SFDN**P**LLSKSDDRIEGF**YVGY**RELSSLEAFTFKTFESLPPTGEPHRVTYELGGLRRSTE
YAVTVQAFNGKGAGPQSE
KQKSPQKAGLRPLYKFPRRDQSLPKRD**VQ**WFEAAPSKILKKKKRNLKEKKAAIKANGQSHIERMIFSLLPCLNACFDFVE
TAAAYPLPKITINRIKRRR
RRSTLPERSFRKKNKKASSFLEPTVITRPGSLISFPFRSWPCGRLTTEPGSSFQGPTSIQPREFCAQKRRRRRTAQQCGR
GRPYARRKDKITAEETNRESFRALVRAPSARDGSNKRAPV
SSGCQKKTTKKRKTEAQLCFARRRQMTCIKCFKVRQRHCWKDGGLFLI
QKALNLIYDSQF
FNIII-5
SCKQKKNQLNFFTLFFLFCPDPPSAPHLRIGGTTSRSVSVNWENEHLQEAPITGYSVYYK
SEGGEWHEVSVPHDRRAFTLSDLRCGTEYLVYIRATNRAGKGPQGETLSVRTNGG
FNIII-6
RPVEPEPGKLFEINSTFVVLHLEAWESGGCPVSYFVVQYRAEGTQSEWTLHSNNVVPQQQLVHLG
DLVPGSWYTLLMSAHNDAGSTEVELSFATLTPAGGRGTSPSLTCFYCYIMDPQVAFYRH

LTVTVPIVSSALVLVIVLGVVCIVL

Is4 contig 922315

Ig1
ARPTRSSEQQGPRFEREPPGLVEFTNSKEASVPCQASGRPAPAVR**W**IKLPDAVTAAEV
PGLRYVRPDGTLVFPKFAPKDLRQDVHSAL**Y**RCV**A**T**N**SV**G**AVA**S**RDVRVRA
Ig2
*VSQPFEVRAYDEFVTRGNAALFRCHLPSFAKDVLVITAWLRDDGLLIHSPITEGESKY*ALLPSGELYIRETDQQDGFRTY
RCQTRHRLTGAVSQSVTVGQLILTG
Ig3 **ambiguity for the beginning of Ig3 see below another proposition**
GPSRARRGPTARTRSHIRAYIRTTDSEVVLPCVAQGFPVPAYQWLRKDEVSGRAEPVPTAGPRISLIGGNLVIRAAVAQD
AGKYYCVVNNTARQDRAETELIVY
Ig3
MVPPRITHLLGQVTALEDSEVVLPCVAQGFPVPAYQWLRKDEVSGRAEPVPTAGPRISLIGGNLVIRAAVAQDAGKYYCV
VNNTARQDRAETELIVY
Ig4
APLRASVKPTRVSASIGHSLRLNCSTEGYPVREVSWTKDSRPLYTSDRIKIIYNEVLVVNGVKRQDRGMYQCFVRNRFET
VQAASEVIIN
Ig5
DEPPVLENIFPESIHKPGGSVSLRCTATGNPLPQVTWDLDGRHLPETIRYRVGDYVTRDNRVVSYVNISSVRTEDGGIYR
CRASNDVGLASHMARVNVYG
Ig6
PPTIRLMGNVTALSGGNLVVHCPVGGYPLTAIRWERDGRTLPSGHRQLVHANGTLVVSEVNRKADEGTYECVAENGRGDI
ARRALHVHVM
Ig7
VGPKVDPFKFPSDLEEGMRSVVVCVVIDGDPPVFIGWLKDGRPLTQDLGAHTEMLNTFTSSLTFHSVGPKHSGNYTCVAR
NPAAAVNRSATMTVK
Ig8
VPPYWRKQPMDKAGILGESVLIDCQADGVPHPQIRWKKMIPGPPVESQTIISNPYIQILENGSLVLREIGLNDAGEYMCQ
AT**N**NVKPSL**S**EVIKLRVH
Ig9
VPAFFKTQFSSQNVRKGEDVRIRCEAYGEKPINITWTKDRQILNFDTETRYKETSTTFPERLVSEILVKATDRRDSSLFT
CMASNAYGRDETNFQIVVQ
FNIII-1
EKPDSPRNLNIKEVTSQSVAMAWMQPYSGNLPLTSYVIQYKKDSEQWTPDVMSARNSPSDLSVVVRNLNPVTTYNFRVLA
ENSLGHGNPSEVVSVITKEEA

FNIII-2
PSNPPTEIQIEPTSSKSIKIKWKAPPSEERRSPVKGYYLGYKLHRSGEQYVYKTLESARNGEIEEFLLSN
LRRNTEYSIRLQAFNSAGSGPASEEIVAKTLEH
FNIII-5
GYVLHYKEDQNDWVKQHVPGTQQSIVLEQLRCGTRYQLYMEAFNDAGKGDPTQVLSVKTEGTA
FNIII-6
PVAPDKASFLVINSTFVLLHLGAWYSGGCRISFFVAQYKPRGESEWTLISNHVQPQTEALLVPQLAPGTWYNLLMTAHND
AGSTDAEFVFATYTETGGTK
RTVPPMVSVNSEDRRFYRHL
GIVVPVACSLVIVLMVALVVCLLY
SRTCCRGRSRVIYETAGEDDRSRMSKTGSRDMVDMVLLSKKLHSSYDETNAKSFYPSPPR
LHQQQQLMLQQQQHQQLATAQNGSQLNNEAQGFDDAGSDCDSVRSNGAGDAAQHRRHQHT
YDVPFHVRRVSVSLLRRRKSLRERVRRG

Is6 contig 682990

The start of this translation is located in Contig ABJB010031034. Nucleotides: 575.198-575.955

Ig1
MFFLCLWSGASSEFRSPHFLHEPPQRVEFLNGT**G**AVVPCVAHGTPAPRVFWMTRAGHPVTEVPGLRHLRTDGSLVLPPFQ
AEDFKEDVHSVVYRCV**A**T**N**SV**G**TIG***S***HDVRVKAGECIGIHSLFRGRSQGLRNELL
Ig2
*IRRRYDVKVYEEFVIKGNTAVLRCHIPEYVREFVTVTAWQVDEA*NLTVENDLFPTGELHIRKVDAADAMSRYQCQTQHRL
TGETVSSPSSRKLTVR
Ig3
ESFAMSPRIVDSRRQVRADKGLSAELPCAAQGEPVPVYQWFRKVRGQAVPLLPGPRLLQLDGTLVAVTDAGLYTCFVNNT
SGSDTVDTELQV
Ig4
ASLSVAVHPRNQKADVGRPASFNCSVTGHPVTSVEWYHNQKPLSRGSSHSPYSVTIPSVRREDRGVYQCYAYNEEESAQA
AAELSLA
Ig5
DDPPILRETFTERTLSPGPSISLKCIAAGRPLPQVTWSLDGLPVPENGRFRMGDYVTSDGSVVSFVNISAVRAEDGGLYR
CSAGNDVGVSEHA
Ig6
ARVNIHGPPFVRRMGNLSVVAGEVLSITCPVGGHPIDSITWEREGLRLPYNHRQKAFPNGTLLVQDVERATDEGLYSCTA
RNKDGLSAQNSVSVRVL
Ig7
VRPAIVPFSFPESLHQGQRFNVLCTVSKGDSPIHIAWYKDDAPVATTGAAAVSVLNVTQFSSTLIFDKLVPEHRGNYTCE
AR**N**QA**G**LVRATSTMVIH
Ig8
VPPRWRIEPSDSIVVKGGTAIIDCQADGFPVPRVRWTKSEGNEPGDYRAISSSSRIHVFENGSLAVHNSDEKDAGFFLCQ
ASNGISPVLSKVVKLSVH
Ig9
VAAHFKSKFKAESVQRGHLVKLKCEAFGDKPVIITWTRDKQPFDPKEDPRYELNETLLSGGIVSEITIRGADRRDSALFT
CLARNSYGTDDTNMQ
FNIII-1 incomplete
LILQEPPDSPPGVKLLEYGSRHVKLSWVTPYS**GNSP**VVKYVLQYREDS
ESATPSLVIESEGTPFYLE
LGVILPASISLIVVLAVGILVYVV
LRKRYSSGSSNSGSSAYGSRKSHLQECLH
LSEVDKSLGKKSMSLEGRLDYYPTPYATTRVTDIDERKLSECSYKQAQDEPLYATVKRTP
RPPRSDIHVYNYP

Is8 contig 825389

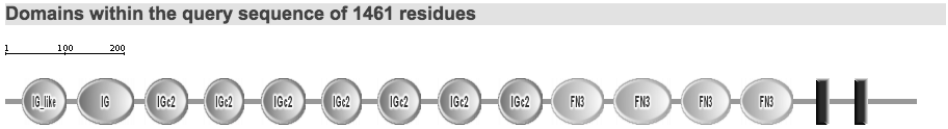
Ig1 uncertain
TLGDLLPGPGSSAPAFLREPPGQLVFPNATGAVVSCSASGDPRPVLSWTNESGSPLGSVPGLRRTRPDGALEFFPFRGED
YRQDVHAAVYRCR**A**S**N**TL**G**SIS**S**RNVHVKAV
(at 1212818)
(MNTGVLRCHVPNYVREYVIVTSWVRSDGFIISVQAIPDENSKYVAFST**G**ELHVRRAGPEDSHHSFQCQTKDTLTGAVTS
SITAGKLVIT)
Ig2

*VQQQYEIRVYDDFVIRMNTGVLRCHVPNYVREYVIVTSWVRSDGFIIS*NSKYVAFSTGELHVRRAGPEDSHHSFQCQTKD
TLTGAVTSSITAGKLVIT
Ig3
EPHSSIPAKVIHWSRQVDGPQGSAVFIPCEAQGHPQPMYRWYRQYGGRLMPQLPMNEPRLVLVGGTLVLRRATVQDSGTY
VCVVSNGAGAEEKNEIQLLVT
Ig4
EPLEVEMRPRVQEVRSGETVTLNCSVSGFPVRSVT**W**TKDSRPVSAGPALRRLVLLNRYALRIQAAQSQDSGLYQCFAGNE
RDSAQGHAYVRVK
Ig5
SEPPVLVSHFEESVVRREEPVSLRCAATGTPLPQITWSVYDVQVHDSGQVRVGDYVSRDGSVISFVNFTKVRLEDGGTYR
CEAANEHGQDSYS
Ig6
ARLNVAGPPTVQPMANRTVVAGRKLLLHCPYSGYPISKVIWRKDGKSLPSSKRVMPYQNGTLALETVSRNDDEGRYSCIV
RNDQDAEATNQLNLRVL
Ig7
VPPSITPFSFPEKPQLGSRASVTCSVPEGDAPIRLSWLRDGVPISSSSSPGVTLGHVDDFISTLVFKSLREEHTAVYTCL
ASNEAALVNYSAPLVVY
Ig8
APPRWRLEPADATVTTGERVVLDCQADGTPEPRVRWKKSAGVQSTEFRTVISSSRMQALVNGSLVIQEIETSDAGGYMCE
AS**N**GV**G**LPLYTVVQVSVH
Ig9
APAKVRQRFLSHMTGKGQTVNLRCDASGDEPIHFFWSKDSRPIKTFSNPRYTIKDSARPGSPSSDFTILLAEKNDTGAIK
CEVSNAYGHDEQITHLSIQ
FNIII-1
DRPDEPPRPEVLNVVSRSVTVL**W**KSPS**DGNSP**IIK**Y**IVQYKRSVDSWEKQLSEMVAEADQSQVTVQDLHPLTEYNFRILA
ENAIGIGPPSEVLTVITDGE
FNIII-2
VPATPPQSVKVSAVGSKKVEVAWKPPPLHLQYGDIQGYYVGYRVHGTTEPYVFKTVTRASGPTTQCIIDNLQRATAYAVV
VQAYNEKGAGPLSDEIMVQTHEH
FNIII-5
DPPPSPIVTVVTRTPDAIELAWTPQEENKDAIEGYVARFRLHEGMDWNEVSLGPDKRGYLFEGLVCGSAYYFSILSYNRN
GRSEPGELLQVKTEGT
FNIII-6
VPQPPSHRTGIVPNVTGLSLALGAWRDGGCPISHFFIQYKSRDDSEWTLLSSRVLPDRDSVAIGDLMPGTWYNIVVMAFN
SAGSTKAEYTTATLTLSG
NPLLEPDKMAETGRESIPRYRSL
IIVPICCSIVVLMAVTVAIMVLL
CRKRTSGTPPSAMDTYGGVRMCEDLKMDSLIMSELEKPGSGDVGGREYYPSPYASSKLPNISRRESGDDGGGPRLDEHGR
VMSAGVSMSPYASSRMVEHTYDVPQHPREGTGLGFFNT

Is9 contig 922315

Ig1
GRALTTPEHLTGPSFSVEPPTRVTFYNST**G**ALVPCTAVGQPRPDVH**W**VRAATGHPVRDVPGVLAARYDGTLVFSPFRAQD
YRQDVHAATYRCL**A**S**N**SA**G**TVG**S**RDVHVRASE
Ig2
*VTTSDVFVIRGNTAALRCEVPASVRDFIHIVYWETDDGLTLHG*DKYQISTDGDLVIDRVDVADARRKYRCITRNALVGET
VSSSGWAQLLVT(MNFPGLKKNLP)
Ig3
DTSNYLPPRIRRLKQTVRLSAGDPLRLACVAQGYPAPSYRWFRKDDSLVLPVATGGGGRVRVFRGFLLIQSTVRQDAGTY
VCAANNSAGEDRTQFEVVVT
Ig4
MSLKVSVSPGTVLTQEGKTVVFNCSVRGFPVSSVSWMKNQQLLVPSNRVRAVGQTVLHISGVQRADRGMYQCVAHGHDSS
AQGAAQLVLE
Ig5
ENPPDFLETFPDQLLKPGSAVSLKCSVTGNPLPQISWFRYGRLLSDRSGLRIGDFVDASGVVTSFVNVSSLATEHGGVYS
CRAENELASVEHTARLSVYG
Ig6
PPFVHRMDNVTHVSGSDARMQCAASGYPITLISWKKDNEGLRPSSRLVSSDNGSLHIVHVSQSDQGWYECAVSNKKGNTA
VGSMFLRVI
Ig7
AKPVINPFLFMKNLQEGMRTTVVCSVLSGEPPVEIDWLKDSAPLSEVHPEAKITRLGDFASSLTMDNVTRRHSGNYSCKA
TSGIATTNYTSRMDVS
Ig8

ASPRWMKQPSDQSSTRGQRTVFDCEADGNPLPVHRWKKNEKLSEFRSVVSSPHMHVLENGSLVIVEVTPKDQGHYLCE**A**S
NGV**G**PAL**S**VAAYLQVN
Ig9
VPPYFHEEFETKTVRSKEDVTISCEVFGETPLTVAWSKDRRYMFSVSRRFVLQEDTTAEGIVSKVFIPSVGREDSGVFVC
EATNSFGKK
FNIII-1
DRTIQLIVQGPPDIPRDIHVDQVTSRSATLFWTQPHTGNSPLLGYTVLYVPEADKVTSAPSSLRTGTSENRATVPGLVPG
TAYILRVVAENAVGKSGPSDEIRVVTEEEAPSGSPYEIRI
FNIII-2
TATSSKTVHVRWKSPLQSTYHGKLKGFHVGYRQLNSRETFQFQTVNVEDDEAKKEPKDNEFEIRGLRRFTQYAVVVQAFN
NKGAGPLSEEATVQTLEF
FNIII-5
DPPSAPQLMITSKSSSTLELEWKFPEVTETPITGYVVHYKSEYGEWQETQVNSKLHKHLLTNLICGNRYQVTITAFNAAG
RGVPSELVNAETTGR
FNIII-6
GPIPPQDKSWSVLMANSSSISVNLDGWSDGGCPITFFVVQYKPHMQPDWVLLSNNIRMAQSPVTIPDLAPGTWYDVMVSA
YNDAGATEVEYRLATLTLSGATVAPLAAQSQESGSSFLRDP
AILVPVACAVVVVLVICLVVGVVLVL
RRRENTYDSCHTQHISTVAPLAAQSQESGSSFLRDPA
ILVPVACAVVVVLVICLVVGVVLVL
RRRENTYDSCHTQHILSAGEMSSGSPSRHLQANMAADYAQSAAGTLQRNRYGNRMHLYDVPLRPKQVPELLCSRT



Is10 contig 922315

Ig1
EVSRPRFVQEPPSRVVFSNSTGAKVPCAVSGYPRPSVTWYSHQGHALAASVGGSDAGPSVVANGLRRVLPDGSLAFRAFS
EREYAPELHHATYRCS**A**T**N**AV**G**TLV**S**RDVKVRAD
Ig2
MEGV*VLEEFEAHVHDDYVPRGNTALFRCHVPSTLRQYLSVTSWTTEDGLVIGRRETHL*Q
Ig3
PSGKSAPRILKAQASVETSPGEDAEVPCLARGHPPPSTRWFRRSSRGLTPVASRPGTVHLPGLLVLRSAVESDQGRYTCL
ANNSVGEDRMDTELLVRL
Ig4 incomplete
NVSVTVSPEEARAELTRPMTFNCTARGFRGGALSFSWLHDGSVP
NNNNNN
Ig5
ETAPELKSVFTKKLVDLGERFSLRCVASGNPLPRVTWALDGGVVGESHRVHYGDFVSSAGDVVSYVNVTSSTRDDGGLYR
CEASNELGSAWHDDRIDV
Ig6
RGPPRVRPMGNLTVTSGTTLVYHCPFTGHPAPKVTWSRGGRDLPHNERQRTFDNGTIIVVDVTRESDEGVYTCKAATPKL
QAKEDLLVKIIKKT
Ig7
VLNPFSFPKTLAEGMQVVITCSVRSGDTPIKIWWLKDGVPFSKTQLNIHEASLGDLGSNLVFNEVGRAHNGRYTCVAEND
GGITNHTAELVVF
Ig8
VPPKWKIEPSDKSSIVGSRVTFDCQADGHPAPLIRWKIALGEDPGKTFKSIISNYHMQMFENGSLIINDVEPKDAGKYLC
E**ATN**GI**G**VGL**S**TVVRLSVH
Ig9
VAAHFKVSYQALRVNKGEQARLVCEAFGERPLAMSWKKDNLILDHRYISSFTQEDTPTADGLTSSLRFAAAERSDSGLYT
CLTSNKFGKDETNIKLLV
FNIII-1
QETPDSPDDIRVVEASSRRITLRWNAPFNGNSDIIGYFIQWKEVAGSWQKDARQLEVSAANTTAVLDDLQPITSYHLRVL
SVNQLGRSDPSSMISVTTDEE
FNIII-2
VPSKPPEELVVVPVTSQILKASWKPPPNFSAHGRIRGYYVGYKPLGSGESFVYKTIDVLDGFVPEISIGNLKRSTKYSVI
VQAFNGKGAG
FNIII-5

PPSPEVTAQTFEHGFVLYWKSESSEWSERGVDGATTTHTLEELNCGTRYHFYVVAFNDVGRSEPSSSVSATTSGGAP
FNIII-6
LAPDKNELVTSNSTAVSLHLRSWKDGGCPIRFFAVQYKLRGQREWTAVPETIDASLAEFYVVTGLQSGSWYHLLVSASND
AGST
EAQFVFATLTLSGATIPPMTLHPEESTAFPR
VTLVIPIICAFVVAFVIGAVVY
MVCNRRTRAHDYSAASQASERACGGDMKGDSISMTSVGKKVYETPRGDPLYFPSPYATTHISVYSGDNDSPSGGPRGHQA
PASAGAGPGGGTPISGRPEHTYDVPFPPKQELLLETASYNPAE**TRYDRL**PRQRFSLYGQKTDQKAVASNERISDEESNQ
DEAESRGFEGNTAPSENMEMSEAECDRDFQIYSKKGRNMSLVQYAKTRPVHSTSYVTYH

Is11 contig 704057 (swaping in fn 1-2)

Ig1
MEFSSESGAVLPCSARGQPTPRITWERKDGSPAAPVDGLRSVRSDGSLVLSSFLASQYRQDVHSATYRCVASNPLGTVKS
RLVHVQG
Ig2
V*VLQKFTANVYDVYVIR**G**NSALLRCYVPPAVKDYVRVTSWVRDDGVTVGTLGST*GIEDRYLMLPTGELLIRDVQSPDTFR
GYRCQVRNVLTGVTDTSATAGKVIVT
Ig3
EPHTQTPPRMAEYRSVVQVEQGDQAFLPCLAQGNPPPTQTWYRLHGPASSSSSTGLGNPTKGLRRSSSPVVPSERLTLLE
GALVLHGARTQDEGKYACVVNNSAGEDRADTDLVVT
Ig4
VPLSAHLEPSVQTVDVGRTANLSCRVAGHPVHGVQWTLNGRPLAKGDPRFTLLSRDLLQVSSVQRDDRGMYQCLAFNQRD
SAQGTAQLVIGE
Ig5
DAPVLEQVFSEQEVRPGTSMSLKCSASGNPLPQVTWTLDGGAVPEVYHIRIGDYVSNERIVHSYVNLTSVRVEDGGRYAC
VARNGVGAAQHSARLNVLGRPL
Ig6
VRPMGNVTALAGRPVTLHCPVAGHPIRSIAWLKDGRSLPQNHRQRTFPNGTLVISDVQRSVDSGWYSCVAQDPDGNSAKR
QVALDVM
Ig7
IPPVVNPFAFPSDLTEGKRAGAACIVSDGDLPISVEWRKDGLPLAPALRASVAEANDYTSFLSFAAVRQSHSGNYTCVAS
NPAASANFTAPMIVQG
Ig8
VPPRWRQEPRDMSAVMGQAVVFDCQADGFPVPVIRWKKAHGRGGRDFSVIISNANVQILENGSLSIREADRKDGGQYMCQ
AINGVGPGISTVVRLDIHGIL
Ig9
AAVTERQLQEYVVDVKRHVEDLVMAAHFERKFQALTVRRGESIALTCSVVGEPPITVTWTRDRHGFNPTLEPSCTRFASR
GLACLEAFLPLVP
FNIII-3-2
GVGPGISTVVRLDIHGILFQPPTEEETHGTVHGYYVGYRVRESKESYAYKTLEASTAAAGHGFTASSSSLHECELTDLRK
NTRYSVVVQAFNAKGAG PSSEEVLAQTLEIDPPNAPSLKLVSSTSSSVHLSWEAAKEQPVSEP
FNIII-3-1
PDKPRGLETTSTTSRAATLVWAPPYS**GNSP**VLKYLLEYKTEPGSWDTDKHLVAVDSTDLSHVVNALKPKSTYEFRLRAEN
ALGVSDYSDSLVLTTDEED
FNIII-3-5
PPNAPSLKLVSSTSSSVHLSWEAAKEQPVSGYVLYQRAEATPGSSSLSSESAGEWSEIQMSADRSAYAFRGLDCGRRYAF
YALAFNAAGRGPQSNTVFAKTEGS
FNIII-3-6
APVAPELQDLVSLNITAVTLQLSS**WKSGGC**PIAYFVVLYKQQAAREWTPAAARLPAPAQQHP
PQSTTLVIGDLSPATWYDLLVTAHNEAGSTEAVYAFATLTLDGE
SPPRLTQAVDSQQRQIR
IIVPVVCVLFVLFMVFAVVCCVV
SRRRLSMARRREDMEEPENTKAVDTVPMSVWEKPDQVACREQLYFPSPYAGSRVCAFVDGVPPPQHTWTTTGRLRAGEHN
EASEEMDAQHQHTYDVPFLRRPPCTEQL

Is12 contig 704057

Ig1
KGGRGPSLVLEPPTAMEFSSETGAVLPCSARGQPAPRITWEKKDGSPASAVPGLRSTRSDGSLVLSSFSSSQYRQDVHSA
TYRCV**A**S**N**SV**G**VVK**S**RLVHVQG
Ig2
*VVLLKFVANAYDVYVIRNNAALLRCHVPPAVKDYVRVTS*WRIENRYLMLPTGELIIREVKTADTFRGYRCQVHNILTGSS
DMSATAGKVIIT

Ig3
EPHTQTPPRMAEYRSVVQVEQGDQAVLPCLAQGNPPPTQTWYRLHGPASSSSSTGLGNPTKGLRRSSSPVVPSERLTLLE
GALVLHGARTQDGGKYACVVNNSAGEDRADTDLLVT
Ig4
VPLSARLEPLVQTVDV**G**RTANLSCRVAGHPVHGVQWTLNGRPLAKGNPRLTLLSRDLLQVSPVQREDRGMYQCLAYNQRD
SAQGTAQLVIG
Ig5
EDAPVLEQVFSEQEVRPGTSTSLKCSASGNPLPQVTWTLDGAPVPEVYHIRIGDYVSNERC
Ig6
LLFPVRPMGNVTALA**G**RPVTLHCPVAGHPIQSIAWLKDGRSLPQNHRQRTFPNGTLVISDVQRSADSGWYSCVAQDPDGN
SAKGQLALDVM
Ig7
IPPVVNPFAFPSDLTE**G**KRAGAACIVSD**GD**L**P**ISVE**WR**K**D**GL**P**LA**PA**LRASV**A**EA**N**D**Y**TSFLSFAAVRQSHS**GNYT**CVA**S**
NP**AASAN**F**T**APMV**V**Q
Ig8
GGDSGGRDFSVIISNANVQILENGSLSIREADRKDGGQYMCQ**AIN**GV**G**PGI**S**TVVRLDVH
Ig9
VAAHFERKFQALTVRR**G**ESIALTCRAVGEPPITVTWTRDRHGFNPTLEPRYVVEEKPGAEGLEYSVHIPTADRRDSSLFS
CYAENAYGRDDTNFQVVVQ
FNIII-1
EPPDKPRSLETTSTTSRAATLVWAPPYS**GNSP**VLKYLLEYKTESGSWGNDGHLVAVESTELSHLVNTLKPKTTYEFRLRA
ENVLGLSDYSDSLVLTTDEEA
FNIII-2
PGGAPRDIKVTPTGSRSLRVAWMPPSESESQGTVQ**GYYVG**YRVRDSKESYAYKTLEAASTSLGSSSSGLQECDLNDLRKN
TRYSVVVQAFNGKGAG
FNIII-5
PSSEEVFSQTLEIGKPRLACNAMPMGADRSAYAFRSLGCGRRYAFYAVAFNAAGRGPRSNTVHAKTDGST
FNIII-6
PVAPEQQDLVTANMTAATLQLS**SWK**SGGCPISFFVVLYKQQAAREWTPAAARVLPEMPQHQSRRQQQQKQQSQPQQAHLP
ATTLVLGDLTPATWYDLLVTAHNEAGSTEKLDFYDNFRN
SALWCLILVFLFLFLSA
KVYSKFQNQRAQQSTLHPKASWFSTWSSSPEDMEEPENTKAVDTVPMSVWEKPDQVACREQLYFPSPYAGSRVCAFVDGV
PPPQHTWTTTGRLRAGEHNEASEEMDAQHQHTYDVPFLRRPPCTEQLVSLS?
EAEYVFATLTLTGDDLDEPENTKAVDTVPMSVWEKPDQVASREQLYYPSPYAGSRASVYADGAQQPQDTWAPTGRLRAGP
LDEGDVQEDEQQADLQTQHTYDVPFLRRPPSSQTQLSSHDGLISSTELLSNHIYSKPAVVYLPPENGKSLRHQHHGSSHL
PVSIIEGYPSGNVSYVYSRPKKKHWSQQDSPYAERKLHKLNSRRYSDEMKQGDMVSRESGLDFAVEAYELSEAECDMPSR
HFPVQR

Is13 contig 973132 (EST ref XM_002400252.1)

Ig1
MDHGDNIALGVPTLADVSASVRRGPFFTLEPPHWVEFSNTS**G**GEVRCEADGDPPPQLL**W**ITVDGSPVTSVAGLRALSEDG
ALTFPAFAADAYRQDIHAAVYRCL**A**S**N**EV**G**AVA**S**RDVHVS
Ig2
AV*VDYKYEPRVYDGFVIRGNTAVLKCHVPSYIRQYTLVDAWIRDDGFTINASGNK*EDRYSLLETGELLVHKTTSEDADRS
YRCRTRHRLTGHLTASSVAGRVTVT
Ig3
DAHAMTHVKMALNFPELKTTVGSHVDLPCVAQGYPPPHYTGRRLSVVESDRMQSSNGVLSIRAVNVHDGGRYVCIARNTV
GEQKIETLLSVA
Ig4
VLLSAEVSPAFQTVAMGLPAVFNCSVEGQPVHSITWRKDGSPLVPDGRIQMVSQSQLRIQTVRRDDAGMYQCVAQNDRDS
CQAAAQLRLDD
Ig5
DISPTLVETFAPQVVKRGDPVSLLCRARGSPAPELTWAIDGDFLYPSHRLKITADRGSLEVRSLLNISEARHEDSGEYSC
MARNDIATEAHSARLEVYG
Ig6
PPFVRPLRNVTVVSGTELALRCPYGGFPVDSLTWQK
Ig7
VAPVIDDHFFPDVIKVEEGTRSRLMCS**V**SK**GDPP**LR**F**RWL**K**N**GL**T**I**GSHGDRSIEATDDSSIIKFARVRFVDRGSYVCFV
SND**AAS**V**N**R**T**VQLV**V**H
Ig8
VSPRWKTEPQNASAVLGASVFLHCASDGFPSPAITWKKGEGNAPRNFSYIHYNFRKHHFINGSLLVREVEESDQGFYLCE
AQ**N**GI**G**PGI**S**KLVFLKVH
Ig9
VPPRFEVKHRSFLLKKGEDFRPQCLAAGDSPLLYSWEKNQNPLDAER*YRVKEEQKQRGVFQSDLLISQATREDSGVFSC
KAINTYGEDTTHFQVIVQ

FNIII-3-1
EPPDAPTGVEVMNFTSRSATLQWNAPYN**GNSQ**ITKYVLQHKLQK
ESWSGPVSQLVVTSSDTTATVRGLQPVTKYALRIVAENALGPGTPSNESLVTTKEEV
EARAGNSEENLREGLLLRNTYLTNLRRLTKYGIVVQAFNAAGTGLAS
DEVIATTLETEYVLHYGTEASDWLQLPLNATKQSFVLDGLKCGTLYRLYMTASNSLGTGE
PGAEVSVRTKGAA

PISPTTDKFITTNSTTATLHLNAWSTGGCPVTRFAIQYRLKFHPTWL
SLADSVNPRRRQYQLTDLVPSRQYQVNVIAHSEAGATQADFEFQTPGAVGGRRMNGYAFR
ALIKPLHDWVRSESTTKKY

Expressed, see EST reference bellow

EST ref XM_002400252.1

Identities = 126/126 (100%)

```
Query  1    PISPTTDKFITTNSTTATLHLNAWSTGGCPVTRFAIQYRLKFHPTWLSLADSVNPRRRQY  60
            PISPTTDKFITTNSTTATLHLNAWSTGGCPVTRFAIQYRLKFHPTWLSLADSVNPRRRQY
Sbjct  391  PISPTTDKFITTNSTTATLHLNAWSTGGCPVTRFAIQYRLKFHPTWLSLADSVNPRRRQY  570

Query  61   QLTDLVPSRQYQVNVIAHSEAGATQADFEFQTPGAVGGRRMNGYAFRALIKPLHDWVRSE  120
            QLTDLVPSRQYQVNVIAHSEAGATQADFEFQTPGAVGGRRMNGYAFRALIKPLHDWVRSE
Sbjct  571  QLTDLVPSRQYQVNVIAHSEAGATQADFEFQTPGAVGGRRMNGYAFRALIKPLHDWVRSE  750

Query  121  STTKKY  126
            STTKKY
Sbjct  751  STTKKY  768
```

Is14 contig 843075

Ig1
MSQGLYRPCEASPQRLGPRFTAPFPAEYRFSNST**G**GWLHCVSQGQPQPRVTWLLADGREAQPLGGLRRALPNGTLHFPPF
RAAQFSQDVHGASYR**C**R**A**T**N**LF**G**TVV**S**TEVRVRG
Ig2
V*VEQYYEVQVYDEFTIAGNTAVLRCHVPSFVKEDVVVVSWEHKLAQKTEVITT*GGRMSVFPSGELHVRRVQPSDASADFR
CRTWHRLTGETKLSSYGRLVVT
Ig3
DLKVNVPPRITNVRSTVVARDGDTVELPCAAQGYPPPKYLWERLPTSDSLSSRRSVLAGSSRFEPSDGSLIIRKVEPEDA
GKYLCLVSNGVGEERATVTLDVQ
Ig4
APLRVSLSPEVLTAHVGHPAVFRCAVSGRPAAEVRWAKDGIPLVIDRARIQLLDERQALRIGSVDTRDGGMYQCAASNAH
ESAQGTAQLILG
Ig5
DTVPVLLESFGDSSVRAGDSVHLKCEATASPAPKITWTLDGTRVHPVRSGRVDLSEATRGEGHLVSYVNISRVKTEDGGL
WQCTASNSAGSVTASAR
VGVYGPPAVRPFPGNRTAVATETLSLHCRLLSYPIDSVHWEKAY

Is15 contig 922315

Ig1
ERRGPTFSSTPPSRVEFLNSTETAIPCEVQGTPSPEIWWARVGAPGPMPDIPGLRHVRQDGALVFSPFRAEDFRQDIHAA
VYRCG**AK**N**N**PV**G**AIV**S**GDVHVRAG
Ig2
*QHFDVQVYDEFVIKGNTGVLRCQIPSFVKEYVTVTSWIRDDGLVIhadsdf*VFPSGELHVRKVDPGTDSHRKYYCQAKHR
LTGKVYRSSTVARLIIIGDGLLSGLAAWFPISQSR
Ig3
DTHVNTSPRLTDRRPVVRARRGDTVKVPCAAQGFPVPSYSWHRVEGGWQVTLESGRVSQADGTLVLRHVAVADAGKYVCV
VNNSIGEDRMETQLQVT
Ig4
PLSATVRPRRTVAVEGSSATFNCSTSGHPVSAVLWLKNGQAVSSRVKMLTRETLHIASVLRDDKGMYQCFALNDYDAAQA
TAELTLGASAPCFQSSFGRMC
Ig5

VPPPPSDRRTVQCTRHARSRDPELVRCARKLRVPESSLEFFRLLGDAKLPQSKRQSVFPNGTLSVLKVERSGDEGSYRCV
ANGPRGDSASGELFVNV
NNNN
Ig7
VAPVVGPFSFPANLKE**G**MRAIVTCSVLEGDSPVRIR*W*LKDRG
Ig8
PPRWKVAPKEKSAVVGENVVVDCQAEGFPPPRIWWEKSSGSRPSEYKVIISNSHIHALENppqGSLMVREAERNDTGFYL
CQ**A**S**N**GV**G**SGI**S**KVIELKVH
Ig9
VSAHFKNAFNSKTLRKGDTAHIKCEVVGEKPLTIAWSKNGQPFSSTIDQRYDIKSTESEESLLSQLEIHAVDRRDSALFS
CLGTNKYGQDETRTQLIVQ
FNIII-1
EPPGAPFNVRTSGITSRSMSVSWDQPYTGNSPISAYKVQVKTGPPVKWKEDIQENVVQGTLTTLTLRGLRPVTTYIVRIR
AENSLGPGEFSQEIQVTTDEEA
FNIII-2
PEGPPLNVQATAVSSSSVKVTWLAPKRDQQNGLLKGYYVGYRQHGSSDSYTYKTLEIAGNFK
EEALLTSLARSTKYTVLVQAFNDKGS
FNIII-5
GPPSEEISLETFESGYYIYIKEQFGTWEEHQISAHQTSHTFQDLQCGSSYQFYVASYNKMGKGEPSEVISVKTQGS
FNIII-6
APVPPKRDALVSVNATRLSVHLNSWSAAGCPIKSFLVQYRLHDEADWVLVSNAVPPDQKVVVVEDLAPGKWYILQVTAHS
EAGSTEQEFTFSTLTRTG
AAIPPLNSLEGQKPAFYRSM
GILVPLVCVVAILVPIVAI
MSFIVSRRRRQAAPNHFRDSCSEDKNLEAMSLSIVKQTGSGLESASPSKDQIYYPSPYALGGREPVLHRQGPSES
DSVHTLKRNRREHIYEVPYPRWSEEEGPYSHITGSAISPTANIYQTPRKSGMKIVL$

Is17 contig 615387

Ig1
RGPYFTLEPPALVEFTNSS**G**AEVRCQADGSPKPSVRWETASGVRASQDGTLTVRPFSAESYRQGVQAAFYRCV**A**A**N**VV**G**S
VA**S**RLVHVLG
Ig2
*LLDERLQARAQDDVVIR**G**SSAVLRCKVGRSQAPYSAFDAWIRDDGYSISRPTYK*ERYSVLQTGELLIHRTNMADTERTYR
CRVRHT
Ig3
IGESASPRMSLFRNVVRVSVGRTVDMPCVVTGFPPANVTNRFLLNFRWFRHQSRKLQTIVDTGGVRQVNGVLTFEEVKQQ
HEGTYICVASNELGEIRAEAGLFVK
Ig4
ETVSLALMPNYQVVEP**G**MSAKLNCTTTTGSVDLSEVTWYKDGRPLKTDVLRVRLETMANLVIRPVEKRDAGMYQCFVGGN
LELAQASAEIAVA
Ig5
ETAPSLTQTFYQRSAKP**G**ESISLQCQSKGRPLPTFSWERDQELLLSDRRVRITSVHISNQVISVLNITRVYAEDSGLYGC
RATNEAGSVAHWAR
Ig6
VGVHGKVFVHQGLSNVTAVPGQDVRIQCRYGGFPVDSVSWYKDDVLLPRNVRHSLDNDGNLRIRDFMGSVDAGDYTCVVK
SRDQEVRATTQLVLV
Ig7
VPPVIDDHFFPETITVDEGSRSRLLCSVSKGDGPLRFQWFKDGQLLSSVPDGSVQYSDDSAMIKFRKVRFRDRGKYTCFA
TNDAAGDNRTTDVVVN
Ig8 incomplete
VSPRIKVAPQNSTTSVGGQVMLDCVAEGFPTPVVTWQKF
Ig9
EPPQFKERFKVLYVRRGETFQAHCSTSSGDAPIAFTWEKNYRPLNCSRCVTRNNSDGSDLTLLGTIRSDSAVYACIARNG
VGEDVTFLQVVVQ
FNIII-1
ESPDAPWGLMLTNHSSRTASLLLHAPYD**GNS**DILKYKVQYKLEQGKYGFGREIVVPAGETTATLTNLHPVSTYEIRVVAE
NAFGASAPSNVTVVTTKEE
FNIII-2
APSGPPVSVSLYTTGSQSLKVTWRPPSRDQHHGVILGYHVGYRVADGAEPGAPSVKQVDSRGANSSHGLETTYLTNLRRL
TKYAVTVQAYNGAGRG
FNIII-5
PSSEEVYATTLETEYVLHYGGDDGDWKSHQLAAHERQFLLQNLRCGSQYRLYVTASNSLGMGEPGEEAVVRTRGSP
FNIII-6
VAPSKEGLIVANKTSALLRLGRWGDGGCPVDRFVLQYRQKLEPAWAAVAQTVAPPPQGEHLLTGLAPGKVYELSVVAHND

AGATP
RAEYDFVTLSPKATKTKTEPMSGSSGGFRFPLQENL
VFIVPALLSALVVLLVLVFLYFYW
RKQAPVADTASEKELPGRKVYAEESFIISELPRKAERSSQDPQGVGGSIFDPRAKRNYHIYTTNQSESTVLSLPAIKHSV
P

Is21 contig 632703

Ig1
GAHWSLAHSLRVSRRRRXXXXXXXTLKSSNMILGFSQTYGQPPASIGWRPVALGSPLEGGGAMVLQMHGQGSIGDPLADV
VGVRRLLPNGTLVLEPFSAQKARFHSGVFQCV**A**S**N**EV**G**TIV**S**RDVHLRG
NNNNN
Ig2
ARYSVLSPTGELLVRNVSSSD**D**GIS**Y**RCQTRHRLTGKAKISDTAGRVI incomplete
Ig3
VNRESLPWRNGTSTRSLSFLIRLAQSIELVALFSGTHCHCLTSKRRWYKLSGGTNESREPLHQGGRFSVSGGTLSIRHAA
VADSGRYLCVANNSLASEPFTVTLTVM
Ig4
APLSAVVVPDEQTVDLGGSATFSCVPSGHPVTSLVWLKDGRTLRQGDPRIQVPLEDSGMYQCLVKNDQDSAQGAARLKLG
Ig5
FSAPTFLSVFSEQSAEPGRGVSLQCSATGSPVPRITWSLDGTSLAADPRVRSGDRVAAPNHVTSFVNISAARTEDGGLYA
CAASNGAGSVEHA
Ig6
ARLNVAGPLRVRPMVPVRAVAGGPLRLDCHYAGHPVDRISWTRGGVHLPSSKRQEVLRNGSLVISEVRQYEDNGTYTCHV
SGPLGQSTSGTVTVNVR
Ig7
VRPTIAPFSFPGGLQAGMRARLGCTVISGDPPFEFDWRKDGRPLSPELGVRAQTDAFSSDLTFASLGPRHNGNYSCVVSN
AAASASHSASLVVQ
Ig8
VRPLWVIEPGDASVLL**G**RDARMDCRADGYPVPTITWERENLYGSSGYSVITSGSDYEIFANGSLLVKNTREQSAGRYLCQ
ATNGIGSGLSKLVHLKVH
Ig9
VGPNFDIKFRSEAVQR**G**GPARLRCEAQGDPPVTLT**W**AKDGQSLGPPATDQRYTFREDPTSSPRRAISILEISSVERRDAA
LFTCRASNAYGGDDLNIKLIVQ
FNIII-1
GLVTQGRGAKPLSMKTTLLWSGELETDSKATSFHFSFNSKTLQRPSANMSVGGGNVWAAVRPLRPAVAYRCQVRAENEVG
IGEPSEAAQVTTGIE
FNIII-2
VPGGPPLEVKATAVDSQTVRVTWKPPERDLWHGELK**GYYVGYRL**DQRGDPYLYKTLQLGSGQEGPHIPEEVLLSPLRKFS
PYVVLVQAFNAAGPGPRSDEVSVSTMDD
FNIII-3
VPSQAPQEVQCAALSSESIRVTWQPPPKDAIHGYLQGYRIWYAQLPASRGEWGCREEKAVTGQETTLVDLRKYANYFIQV
AAFTQRGLGTESEPVFCRTL
FNIII-4
EDVPDSPEDVKVLIVSATSLLVAWKPPVHRNGLITMYSIYAKTLDKRVRTELPFIPLLLSHTPLEYNLTLVPRNARVEVW
VTASSRVGEG
Ig10
MDFDEVVRVAP**G**EDVHLA**C**RFLGTAPVH**W**DWKHGYAQPPSGDGAVQVRPDELGADGSLALRRIEAA**D**A**G**N**Y**TCNVRNKLA
TDRRHVALIVRGQHR
NNNN room for FNIII-5
FNIII-6
APVAPNKEDLVHVLNGTHVKVTPSA**WRSGGCPLT**RLSAEQRLQSLTDWTSVWNHTSSSSGALPQDLPDVLLGPLQ**PETWY**
VVRLMAANAAGVTWVKHDLVTLGAQX

Is23 contig 672165

Ig1
MWIRECASLFQQDRRGPTFLYEPPRRVSFSNAT**G**ATIPCSAVGTPDPRVTWTSADGAPVDDVRGLRYARPNGSLVFPPFR
AEDYRQDVHATV**Y**RCA**AA**N**N**AV**G**SIV**S**RDVAVRAGESS
Ig2 not found
Ig3 second part?
MLAGSSKYHTFPEGELYIRDVDKSLSYRSYRCQTKDKLTGESTRSSLPGRLIITGESPHSN
Ig4
VPPRMAHSRQRVTATIGDTATLPCAAQGSPPPQYRWYRDDGSPVFLDQRTSQVDGVLVVRKATLRDAGKFTCVANNSAGD

DRASSELVIT
Ig5
EPLTATIQPPRQQVHVGQTAIIKCAVSGHPVAAIVWRFNQRPLPISDRVSVPSADTVHIRSVKKEDKGM**Y**QCFVHNEVDA
VQAGFELSLAGKL
Ig6
DLPEFQDTFRPETVHPGTRFSLKCSASGNPLPQITWSLDESAVPETHRVRFGDYVTQAGVVVSYLNFSVVQVEDGGDYRC
TANNGVGTVLHTARINVP
Ig7
VKPTIEPFSYPSSLREGQRSSVMCTVISGDLPINITWFKDDQPITASNPGTAGILVNTVSDYSSTLLFKSLRLDYRGNYT
CVAANEAGTVSHSAVMIIH
Ig8
VPPQWIIEPSETSVVKGRSAVIDCEADGFPMPRIRWTKAEGDAARDFKPVVSSAHVQVFENGSLAINDAKEEDAGFFLCQ
ASNGIGQGLSKVVK
Ig9
FAHFKSKFSAEMIRKGQNTRLKCDATGDKPMRIAWMKDKLVVNPKQDPRYELVETIQTTGVTSEILIRQTDRRDSALFTC
VATNNFGHDDTNIQLIVQ
FNIII-1 first exon not conserved
GLFDTGRSSDQSDRYEEMLEAGRSDLDLVATSSSDDDEIVTDELTVVRTPSHVYAGWKRKTQGAKWHAKMINLSTSATET
SGTVRGLKPALVYHFRVYAENRIGRSDASHSVKVTTSEEA
FNIII-2
PGGPPTKVRAQPTSSRSLKITWNAPNKELHFGVIQGYYIGYRVAATSEPYIYKTLESEMDAGEGCVLTGLSRFTQYSVIV
QAYNKKGAGPPSDEVVVQTLDS
FNIII-5
PPSAPYLHAEATSFTSVSIKWERQSSDQNPVAGYVVRHKESGGSGDWHETRVQGDQNALTIGDLKCGSAYQFTVRGYNAA
GAGDTSDVLTVKTSGA
FNIII-6
APVPPDRQSLLHYNATRAVVQLSTWHSGGCPIQQFTVKYRRQKDLEWTTLQTGLLRDKRLEIRDLSPGTWYTLQMTAHNS
AGVTEAEYAFATLSKHGADQVTPRTEVHRETSSVVSDAT
VVIPVVVSILVVIALLVVV
CMVVRKKHSSGSSQSGTYANGSSLYGTRKNGMQEAMQMTDLEGKVGKECSTSAFFPAP
YATTHLGTRGPEKRADHQDEPLYATVKRTPRPPSTTISEPHYRTFREMYFDLVLPLLVLG
VLCDGEEAKEPKIEPVEDNNVMTLAELAHIAENFDDLDAFKDRWVLSEATKDAAADSVAK
YDGNWQVEAAALNHLRGDVGLVLKIQCFSAQDQGAPPRHRR

Is25 contig 634467

Ig?
MRTIISNPPRLRQAIPGPPLAFSQPCRLSSPTDSSECVRTGCVGPCRTSRARGWGRRVRRSLLPDAESFGDSSGQATASS
ASEEFAYPP
Ig?
VKSVRVRDVSSVTALK**G**KSVSLVCPLYVAAWASVSWEKGTISSNPPRLRQAIPGPPFAFSQPCRLSSPTDSSECVRTGCV
GPCRTSRARGWGRRVRRSLLPDAESFGDSSGQATASSASEEFAY
Ig6
PPVKSVRVRDVSSVTALKGKSVSLVCPLYVAAWASVSWEKGSSKIPFNHRQRVQPDGSLSISNVQQVSDDGSYVCRFTDS
RNQKHTGNVLLKVI
Ig7
EPPVISHYEFRQDMQVGMRIKVFCTVVRGDAPFLFTWLKDGVPVDPAAGVQAGLSVQNQRDYSMLSADSLQLEHSGNYTC
VVKNQAATTTYSAMLRVN
Ig8
EPPKWEVEPENAAVVQGRNVQLQCSANGTPQPTITWMIASDSTREEFLPLYNSHKYGLFPNGTLSIHQLEPEESGYYLCK
ASNGFGEDLSKLVFLTVK
Ig9
RPPKFDVKFRAHAVKRGEKAKLACTATGDLPIAVSWSKNNDRVPDKSKVSTVANQSVSSVTSTLVVSTETVEDSGIYSCM
AKNHYGSDETSMRLLVQ
FNIII-1
EVPGAPVNVTVANATGNSLLLSWAEPFR**GNSA**ITRYLVQFREAGSDDEAALRNLTTNTSLTLASIGSLRPARVFSLRVKA
ENGVGWGRFSGWVTANTEED
FNIII-2
SPASPPVNITARPTGPNSIKISWEPPKEEDWNGHLKG**YYISY**RPVGSSDQYYHKTVDVHNPHQRQEIHLTNLRLSMSYSV
TIQAFTSKGAGPMSQEVLVKTLDD
FNIII-5
VPPSPPTLEVVSVTTSSVTLGWSLKTSFGNPVTEYVLHQRKDSDHWQETPISTVQPLHTVRDLECGTTYQFYMTAHNSLG
RSEPSDVIRAKTDGAA
FNIII-6

PLSPSKEEFIQAAQRHATLSLRSWKSGGCELLDFSVRLRQGGPPQAWANLAEGLPANQSQFLLRNLTPGATYHVHVVAKS
TAGATEAQYEFATLNGTT
HVASVEATSTQPKRSTLPSMTD
LEIIVPILVSSFVVLVVIIVGCILC
SRESLCAERDNCARPELRSNYSEEVVAMKELANAAECMARCEDGMHAPQMGSPFPTAQSIYAQRPGKSLTRTKPRERPYE
SLMVNMNPYPADGTT
TSTLSRKEHEDVQV

Is26 contig 780014

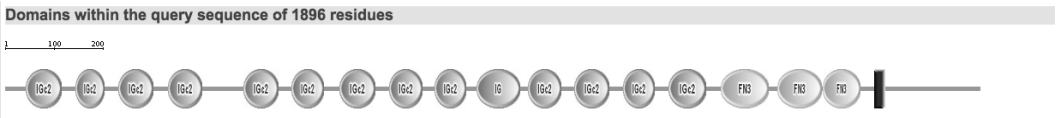
Ig1
MGTLGRSEAIEGPRWVTEPPARLLFSNWT**G**ATVRCSAEGEPRPEVW**W**VTSSDGANVTTLPAARAQLVSANDEQLSFAPFR
DHQFKADVHRAAFRCKAHSARGTILSTIVQVTA
No Ig 2
Ig3
SQTPSVTFHSGHVTVDKGSSADLVCLAQGSPPPKFKRWYKRQGQRLLPVATTPSTSASPTQMDGVLHWSGSVQLDDAGQY
VCVASNNFGEARASLQLSVH
Ig4
ELSAALRPILVRAEA**G**DSVAFQCNTSSSLPDNDVSLDWTLNGSPLPLGFERLERGFVRVSSVARHQGGMLQCFVSSRDGR
RSAQATAELVVG
Ig5
ERAPRLEQTFETPGAVSPKSSASLGCRVSGDPPPSVSWTLDSAWPIVSGGPRLRLWSTSDGVTGDVISFLNWTSVEAGDS
GQYVCRATNAAGRVQHAFRLNVR
Ig6
APLFVRPAYNETALVGATTRLQCPFGGYPFDRVVWYKDGSELPVNQRQSVFPNGTLLLETVDKAKDQGEYTCSVDSGTGT
TVQQTVRVIVRT
Ig7
GPQITPFRWLDELQEGMRAGLSCFVHSGDAPISLEWLKDGLPLRHAHVHSPQGGFMSALSLASLTPQDDGNYTCRASNAW
ASASYSAVLR
Ig8
VKVAPTWRTEPKDVVAVTGHSVVVDCQAHGEPPPHIRWKTWEPGPYRAMVSSSRVHILVNGSLSVRSIETR**D**A**G**LYLCEA
SNGVGAELSKVVRITVR
Ig9
MRQHRRLPMPHPPPLLSHRQNRTGDSMTLLWRYPDTTFLDAVNAPFGDASHLRENKARLFFTCAENGSELPVNQRQSVFP
NGTLLLETVDKAKDQGEYTCSVDSGTGTTVQQTVRVIVRSKDGATSLPFSTLRIVQTK
FNIII-1partial
NVPDVPADVEVGEASSRYVRLSWIEPFG**GNLP**ITQYLLRWTNKEGSWEDSVSVSGTETKVTPDVPADVEVGEASSRYVRL
SWIEPFGGNLPITQYLLRWTNKEGGHKRRAKRCRLESYRKNGVSLDEGPKGRNLHGGTKELGHRNSLILVSGAELANRNR
GHVPVTAMLGIIVGFFQPDFQAQATGVIKEVGTNIDSAEIYEKGKYAGLLVVASSWEDSVSVSGTETKVTVRGLEPSTSY
LFQLRAENRLGAG
FNIII-2
RNAPTNVQLTAVDSRTFEVKFED
DVSGAGRVDGYYVAYRRDGSPEPLRYQTLHERVGVVSGLDRDTLYEVQVQAYNAKGPGPPSRTHAVRTLVA
FNIII-5 partial
AYYILWRVDGAAEEQWREQSVGSDRNGFALSGLACGTRHQLRMRAASDVGRGPEGHLLTASTEGG
FNIII-6
RPVLQQPDRLVEANSTAAWLRLDAWWNGGCPISHFAVHYRSAASGDADWTLVSSHVPTRLDEPVVLVDLTPGSWYVLLMV
AHNDAGTTRSQVNFATLTPSG
DVPSQKSHLLNSKMASFYRHL
TVTLPIGSSVLVLVVVLAVLW**C**VLH
RHAEDAAARGTPQGKM

Ix27 Ix28 Ix29 three genes on the same contig 650268

Is27

Ig7.1
QQTGRTFRLQPFSFPTDAIEGNKVSTVCATVTGGISSGVEFTWFKDGRKLVQDDRIRVRSFPDMSTLVVDSLRQGDSGNY
TCVGKLRNHKDSHTETLRVL
Ig8.1
VPAKWIHEPADVSLKETGNSTVICEATGVPRPTIKWTKEGIALSTQSPSLVFLKATKSDAGNYRCTADNGLQNPLTKQIQ
VTVF
Ig7.2

VEFRLQPFHFPTDAVEGKTVTVTCTTTTAVSGVEYRWLKNNKRVSESAKMRLRTFPELSSLIVGPLEASDSGNYTCQATY
NGKKDSFSDTLNVLGK
Ig8.2
VLPSWIQEPEDIKLMEGSNLTLPCRAKGKPQPNVVITKEGNYARATSKELDAAALSDTLDISKSTKHHSGTYVCKADNGL
GHPLLK
Ig7.3
VLRLQPFYFPSASKVVEGTTVTVTCTTTSGITNVHFRWLKDGREVVDKAKVKIIQHSLLSTLVIGQVDRGDSGNYTCVGN
IGEKLDSHSEVLS
Ig8.3
VLAPPEWVVEPEDIKLHQGGNGTITCEATGNPTPTVKWRVRSQNGAAKETSASGRNLLQLPNASKSDAGTYECSAVNGVP
EDIYKRV
Ig7.4
NALVVKVQPFSFPNDLLEGSRVSVTCSLRKVSSDARFRWLKDGKALDGNRYRRLSVRTEADFSMVTIEPTRQEDSGNYTC
MVTSKGRSDSYTAALVVF
Ig8.4 (broken in two, might not be coding)
ASPEWTESPQDVLVTEGGNASITCKARGNPAPDVTIRKASGAQPSLVQGKSAGTLQLTKTSKHDAGNYSCTATNGFGTAI
EKTFLVKV
Ig7.5
VQVAPFYFPEKTVVGDTIKIICYTNTEQTPLSFAWMKDSKPLRVGDTVRIKTQPDQSAITLGPATASHSGNYTCRASTAK
SSSAYSAQLNVF
Ig8.5
APPAWIQEPSDRRVVKGANLSLPCAASGHPEPKLTWYRITRPTGIFFLFRKNSDGSIFFVRVQKESQGMYRCSAANGIGA
PLNKTVKVTVTD
Ig7.6
VPEIQPMTFPSNLKEGARFRATCSVITGSPPFTFRWLKNNKDLQEDGAVTIENWKDYSNLAITKLAKSHAANYTCIATNA
AGSDRYTNGLVVN
Ig8.6
APPRWLLEPHDAVVLMGGTVRIGCQSVGYPSPVITWEKYGASGGVTVGGTVNGSLEIYNASKRDGGTYGCRASNNHGELI
EKRVTVKVI
Ig9
VPARFEEKFKVQTVRRGEGATLQCTALGDTPLEISWSQEKKPLAFAPVTRYEKFESTTEQGVSSELLIPTTDRSDAALYT
CVAKNEYGSDERNIKLLVHHSSYPAENKLL
FNIII-1
PKAKKVPAQPLDLRILEVWNRKVNVM**WS**EPYS**GNSPV**TNYVVHYWRDKGESPLVLCESQGPHRLHEETVSSTQTSAVIAE
LHPGTSYSMTITAENEVGQGPPSDPIRFQTSEEEP
FNIII-2
GGSPTDVWAAAKGPTSVAVSWKPPPRDTWNGELK**GYYIGY**RSAESNQPYSFKTVETVTNDTQEVTLVGLSKSSRYSVIVR
AYNAIGKG
FNIII-5
VQDKVGPRKSNRSSNHRVTLSLFQGYTLHYKKGSGPWHHIPVVASDDTSYTLTDLDGGATYRVYLTASNQYGRGSGSEAI
IINTVGQG
SDKPWVFYKDPSLLVPVASF
LVALLVVIGVVSVCI
KKTKAHKNLERSALEAEKRQSYAGDAQQRYIDVQEKRAYLSAIPTHEKTIVVSVSVPTGESQKNRRCRIRSNCVLRRERP
ATVVPTQKQRLVAPLRQRRITGDRGCLAAQGWIEGAHISSDYISGAAIRGSTAGVKTSPAVVMATAGYPRCHVIHASWPR
KRKLGGHTQLSIGSPNTRLLTPIC

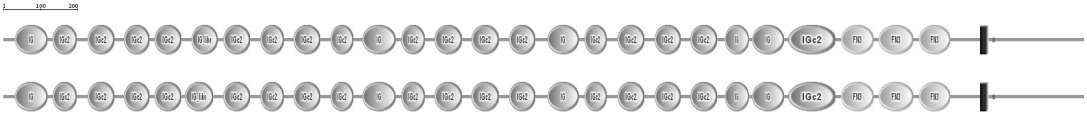


Is28

Ig7.1
*VRPQDAISPRVREPCSSKGGFVSNCYEDEVEVLRAYVHNFAALSILEE*LRNLLRWLKNGRRLTDEKKSVTVTDNADFSVL
KLPSLSLESSGNYTCI**V**S**N**LF**G**SASHSATLR
Ig8.1
VHASPHWVQEPRDVVVTSGERVHVPCQASGYPEPTILWTRKQPGRSEELASSENGSLVITWARKEHEDLYTCKASNGIGQ
RLEKTVQVAVKFS
Ig.7.2
PKITPFSFASKLVSGQRATVTCSTFEGDRPLTFAWLKDGSTLSKHNNVEWNEEKGYSTLNISPLSLQDSGNYTCVV**SN**S**A**
GSDSLSSSLV
Ig8.2 daphnia

VHAPPRWIQAPTNQVVTVGDTAMMVCSASGFPLPTIRWSNRGHPLREDNARVRQWHNGTLVIARTTKDDGGRYRCQAGNA
FGDILEEEV
Ig7.3 daphnia
VPPKVLPFVIPKLLVGERISITCTAASGSKPLTFMWLKNDSALRGGSAVHIADSSDYSMLHIDNLKNDHAGNYTCVVSNA
GGTVSYSDTLHV
Ig8.3)
*MNATALYACSRHSTLRKGFHIVNDSCLLNANGHSLREDNA*RVRQWDNGTLLIARTTKEDAGRYICQAGNTFGDMLEEEVL
LT
Ig7.4
SIPPKVLPLVIPKNLLVGERMSITCAVASGSKPLTFVWLRNDSALRGGTSVHIADSSDYSTLHIENLKIEHAGNYTCVVS
NAGGTVSYSDTLHVKGK
Ig8.4
APPSWTTEPKDVTVTAGDAVMLECTGTGFPKPTISWTKVGKNETSTNTADGLFKIATATKENEGHYRCDITNGIGGSLTK
TVSVA
Ig7.5
VQTKILPFAFPKSLLIGERVSIICTTTAGAKPLSFTWLKGGKPLTKGGDVNIANSPEFSTLSIENLKLTDAGNYTCTVSS
SAETVSYTDTLQ
Ig8.5
VKAPPVWLTEPKDTYVIAGHQVTIPCKGEGFPPPSTAWTKLGKEITRLDGSTITISSAVKSDEGAYRCRIDNGIGTALEK
TVHLAV
Ig7.6
GILIAVPPKIQPFAFPKTGTVGERSSVTCTTIAGDKPFKFVWLKDGLTLRQEGNVKIVSSSEFSVFNIEKLSLENAGNYT
CVVSNAGGTVSYASTLEIK
Ig8.6
APPTWKTEPRDMSVTAGQKVSITCDGNGHPQPSVRWTKEGDRGSSDYRTKTIELPSASKQDQGSYTCEIANGIGEAIRKT
ITILVK
Ig7.7
PPSIPPFQFPKNLQVGQRISVTCTISVGDTPIQFAWL**K**D**G**AALSTASPNIRIVDNAEFSTLNIAPLTLDSAGNYTCSV**SN**
K**A**GYTSY**T**APL**V**
Ig8.7 incomplete
ATWNGRFVLFDTVPAYPITSNIYIKYVICLSAPPRWTNEPQDITATAGSN
NNNN
Ig7.8
PKLQPFHFPQGRTRVGESASALCALVAGSPVVRFKWFKENVAIDGKLPNVNVKNDKRVSVLTIESVTLSSAGNYTCIADN
DYGSDANSAALVVEG
Ig8.9
APPGWKKEPRDLSVSAGQALQLECSATGYPLPKVTWKKDGENPKNEQTLIASQDGSATLSVTESTKETEGRYFCEADNGV
GAALKTALFIKVKR
Ig7.9
LNDFSEAPKIQPFTFNDKVRIGGRAVGSCIVVTAAAPLTFTWIKDGVQLRDKTGLSIQNNRLVSLLIIETADLSSHGNYT
CRASNVVGTDAYTAELK
Ig8.9
VEAPPTWKHEPQDVSAIVGTNITVECRANGSPIPQITWTKSKSGVPMQKDNLIIQNIQDTDAGSYTCKAENGVGPSLHKT
IRISIRA
Ig7.10
AEIPKVHPFSFLKTLSEGQSALVTCTVTEGSKPVQLQWLKDGNEVRSTGTVKITRQETFVALAIEPVQIEDSGNYTCVAK
NRFGYDRYTSLLEVH
Ig8.10
APPKWTQEPVDVTLTSGETAVLYCGATGHPTPAIKWSKLGADLKGTAKEELQVLANGTLLLSSSAPEDTGQYSCQASNGI
GSPLTKTISLIV
Ig7.11
EAPKIQPFQLPSRVKAGEKISATCNLVSGTPPVTFEWLKDGSDVTGLSKDVSYDGNLISVLAITSASLEAQGNYTCRARN
HFGSDSHTVQLKV
Ig8.11
EAPPVWTKEPEDTVGTIGGMLNLTCSASGSPEPAVAWKKLSVVSLFPLAEVHRGTYRCEANNGIGGTLSKTITVSVR
Ig7.12
DAPVIQPFVPPTDVVTGLATKIFCSVQKGSRPLTFSWMKDGRVIRNGVTSLEDYSTLTMDPVTAQSAGNYTCVVSNSAGT
DRYTSTLEVK
Ig8.12
VPAKFAEKHSVVTARRGENARLVCDAQGDQPLTVTWSKGATKIDRAGSTRGNATKGTEGAHGARDGEKENPADQYYYACA
SGSLSRSLSFSLLPFYSLFPRNG Ig8.12
Ig9 partial
MTESGLRSELFISSTERSDGAVYTCRADNEFGRDERTSKLLV
FNIII-1
EVPGQPQDVKVSETWTRSASVTWSPPYS**GNSP**VAKYVIQFWKDSGAAHRLQEVAVPGSQTSALVGDLHPGSTYQLNILAE

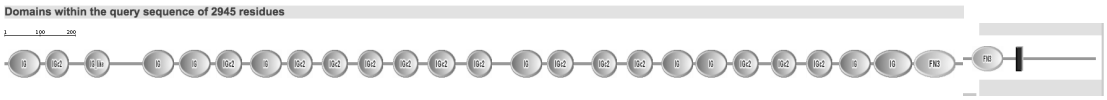
NSVGVGQASTPVKLHTGEE
FNIII-2
EPSAPPTDFHVEARGPSTARVSWKPPPPDEWNGDLLGYYIGYKPTSSGQPYSFRTSEFKPNTSHEFFLTGLQRGTEYSVV
VKAYNAAGSGVASHELHVKTLDGDVPPPPKVFVSGTSHSSITVTWHQQF
FNIII-5
TGVRGFVLHYRAEDGLQDWKEVNVDARTSSYTVPRLESGVLYQLYVSTTNEYGMGDPSEIITVRTHKNGSEGPSFTVRGS
REPPPFQPSSAKKRQHGRGTMSVVGIRFVMPDMQSPIFGDASTP
LYLNLFIMIPVLASLVTIVLVVIV
TCVCLQRIKRRPNQPPGPPPGTMDRRSKQYAAAMEGQPQSKRCSSVFSLSWCIQRPASILARRSNYERRFEELVLDRAQC
LPKTRHVDWILYVIFLDRANTPILSFHIRSRRSRRY
STLRGLQKLAVTSDFRESKRSRFLKVMHGRSSKLIAVIYTRGKVLKVCTQRRLSLYRHLKDGRVPKHIFMLVICRARSIN
CRNGLFRSLCKQKVDVLNFFREQVHWTVSIPATPAAGWHLENKPNISLSCESSAVTSTTARHAEKAQQAGEYTGLSQRYV
EVQPPPLPADHPCALYPAPCATLPMTEELEAKMARHNVNQEMKTFLAQLSPRTHAFCNAIRNVRSRVERCAPKLANAARD
GTTPKLDSRSNRKRGLSTTNERQINNLDDSKASSHGIRSLVDIEIGGLYAYSAETKLTEKKFSVLSVRMSCNCIRIKTVG
GIAGRL



Is29

Ig7.1
AEPPKLNPFSFFKRWQTGEKTSVTCMVTSGTPPLKFVWMKDGKELSEQSSNLRMKHEPGYSMLFIEPVELLSGGNYTCVV
KNRAGLDSYTTFLD
Ig8.1
VEAPPKWVKTIGDGKIAYGSEAKLQCQASGSPVPTVRWQRFDEASQTWWTVNVAGESLVIPRVTLNETGRYKCSADNGVS
PSLEHTLTVAVYERCRSFVIY
Ig7.2
EPPKLNPFSFFKRWQTGEKTSVICMATSGTPPLKFVWMKDGKELSEQSSNLRMKHEPGYSMLFIEPVELLSGGNYTCVVK
NRAGLDSYTTFLDVE
Ig8.2
VEAPPTWKSVPKNVDVVEGEPLVVSCHAHGSPTPKVTWMFKKEGDDRWGHFADAGASTRTLDNGTLVLSGTQESQTGSFK
CVADNGIGTSIAHHFSIKIR
Ig7.3
VEAPKIQPFSFPHRLKVQSKTSVTCIATDGTPPFAFSWLKDGVEVTNMKNIRREKKENDYSVLIIEPVEATNAGNYTCIV
KNKAGFDSHTTYLE
Ig7.4
AVSPKIIAFHFRKTIKPGENARTTCLVEAGDAPMTFSWLRNGVAASLTRNVQIQSHADYSILNVNPVDATSAGNFTCIVK
NKAGFDSFTAYLDVE
Ig7.5
VAPKVQPFQFRKTTKPGETVKTTCFAEAGDPPLTFSWLRNGLDVSSLKNVQIKSHAEVSLLTISPVDASSAGNFTCIVKN
RAGFDSFTSLLE
Ig8.3
VEAPPEWKREPADKTGVLGSNVDIDCWGTGSPAPKITWHHVKAHNERPIDEIFQSRAVTYLNGTLRLHELQVGDSGSYTC
TADNGVPPVLKKTV
Ig7.6
VNSAKVTPKLHPFSFPGTAKPGNNARTTCFLAEGDTPVTFSWLRDGVDASTLKNVHVQSQTDFSVLSINPVDARSSGNFT
CIAKNRAGFDSFTAYLD
NNN
Ig7.7
VPKVQPFVFPPAVKPGSRVSAVCSTTSGGSQVTLSWLKDGQDIGSTKNVFVDTKRGASIIIVEPVEISNAGNYTCIAKNR
AGFDSFTAPLDV
Ig8.4
APPSWKVKPEDKRVNIGDRAVIECLATGSPTPKIKWKKLRKTSEKDIQAEWADVESSSFVKIHQNGTFVLEEVSTADAGQ
YACDADNGMAPSATLVFSVAVN
Ig7.8
VPKLQPFIFPPTVKPGSRVSAMCSTTSGGSQVTFSWLKDGREIANAKNVLVDTKRGASFIIVEPVEISNAGNYTCIAKNR
AGFDSFTAYLDVQGV
NNN
Ig8.5
APPLWKKRPEDVRVNIGHRAIIECLATGSPTPKIKWRKQQKEGKKARIAYPWPLAWPHLKMHDNGTLILEEVSAADGGQY
SCEADNGIAPSATLAFSV Ig7.9

VRAGVPKLQPFIFPTNVKPGSRVSTMCSTTSGGSQVTLSWLKDGKDIANVKNVLVDTKRGTSVIIVEPVEVSNAGNYTCI
AKNREGFDSFTVSLIg8.6
APSWKKKAEDVRVNSGAKARIECLATGSPTPRIKWRKQAKKSGWADLESSNLIKPYENGTLVIEDVSTAEGGQYSCEADN
GMAPSATL
Ig7.10
VPKVQPFMFPPTVKPGSRVSAVCSTTSGGSQVTLSWLKDGKDIGNTRNVVVDTKRVLSNILIEPVEISNAGNYTCI**A**K**N**R
A**G**FDSFTAFLD
Ig8.7
APPSWKTKVEDVRVNIGDRAVIECIATGSPAPRIRWKKRDIEARWIDLISSGAVRANDNGTLILEDVTTTDAGQYLCEAD
NGVAPTATLTFSISVNV
Ig7.11
VSAEVPRLQPFTFPSDVKPGSRISTHCLTSSGGSEVALSWLKDGRDVGDTKNVFVETNKGLSTIRIDPVDISNAGNYTCI
AKNRAGFDSFTAILD
NNNNNN
Ig7.12
VAPKLQPFHFRKTTKPGDIVKTTCVAEAGDPPLTFSWLRNGLDISSLKNIQVKTHGDVSLLTITPVDAASAGNFTCIVKN
RAGFDSFTSLLEVE
Ig8.8 (two exons)
IAPPFFKKTSPDTDVVQGNSVTLTCHATGSPQPRIEWTRTIGDSDKPEDVRRSHRAQSLPNGTLIIEDVADEDEGKYTCM
ANNGIGTVSHSLFMHVRG
Ig7.13
VAPSIQIFASSEVKAGDKVTATCVLKTGSQPLVFLWLKDGKEVSSLPNVKVKSAEDFSFLIINPADVHSSGQYTCVVKNA
DGTDSRTVQID
Ig8.9
APPQWQRVAGDTEVGLGATKSFECIALGSPKPRVTWSKRTESPNGWSPLHGLSRVSMEGDRMTLMDIEASDSGSYSCEAS
NGIGNPLRSIFRL
Ig7.14
RPIITPFSFPTDLSEGVSVQVLCAISKGTLPVYFTWLKDGKTLRETRAKITTADKFSVVQIDAVAPVDVGNYTCF**A**K**N**LQ
GTDSHSAMLE
Ig9
VPARFEEKAAVVTARRTEVTRMKCQATGDQPLSISWAKGSVKLDKRTSARRCRGKTFLTHARRYEVFETLTTDGLLSELV
IRDTDRSDGALYTCNAENKYGKDDRKVKLIVQ
FNIII-3
EVPGPPQDVRIRDVWSRSASVSWSASYNGNSPISKYIVQYWRDHAAWSSALFCGIRKLRARSKIRYRNLRQSCHGNGTFY
QSYREKKVMAY
MVRELLPGTAYVLNLVAENAIGRG
FNIII-3
ESSRTVVFHTGEEESTNCRPQRTFPKFGIVNFVPQNDALSTLKKYWEAPPREHWNGNLQGYYIGYRPRDDADSPFSFRRV
EASSNVSHEYLLGGLQRGTEYALVLRAYNSAGSGPASQEKTVKTLDGG
YESTAMENQDSDDGVPLYLDMA
LIIPAAAICLAVLVILISAC
ICVRKMKSTPRPVPEILRYDPSSLNTETMMSQRYVEMEKMSDNDVVMVAPYDGSTMRNGTELRGTSDRQEMKTYVPKPS
TLNHQKSQLKPVQGDRARTESDAILLSCTPKTNLGEVVEQGRGQERDEMMRNVRMWTPLSSGLRNEGGGMLLRGIRVLRA
RGVPSLESGGGCCRSKMADVQGIAEKREPRFTARVNDLSIGCRKLKIRENVT



Is32 contig 26264

Ig9
PARFEQKFSVESVRRGDTAILRCEAVGDSPMGVTWHRNDDPLPLDSPRLQVFESVTDRGTASELHVQGAERS**D**N**G**L**F**SCL
AKNGFGSDRRSIKLVVL
FNIII-1
EVPASPLDVKVDQSWSRSANVRWNAPYS**GNSPV**SKYIVQYWKDHGERATLEEASVTAPQTSTLLRDLQPGTSYIVRALAE
NTVGRGSPSESQKFQTKEE
FNIII-2
EPGGVPTDVAAEPRGPSSLRIKWKPPPKEQWNGQLL**GFYIGYR**PKSSEDPYSYQSAPMTDQAEEEHLLAGLKRATEYAIV
VKAFNAAGSG
PGSQDIVARTADSDYILSYREETGPWRELTVPQADNSKYSLTGLREATRYQIYLQAAGEGSTSAPSEIITVLTEGGALSD
ASMPAPQGSQSRELPVYFRLS
VVAPAAASLTIVVLVIAGACLFV
SHERRKYQNVAVPPLKPLKTGTCMSGPGSLGRPSGQRYVDVDQRFGPPQPRTPRHTDGVDRGGARPLLALRQHHRGGGG

DSPLRHRGLTKSEGDLNSAMKLSEKVGKNEMVDDITDQVTEEAKRACDREGRPGGPPGLELNSAVYKADDPAVADPNSSQ
PNNMAVAFELNL

Is35 contig 682990

Ig1?
KYSVLPTGELYIRNAGPSDRLGSYHCKTKHRLTGEVATSASSGRLIIQ
Ig2
APQGAVAPRMTDTHPVVLAVEGQDIVELACAAQGFPVPSYRWYRELDGRLSDLTRDPRTAQVEGSLFLSGLEVKDSGKYF
CLVNNTVGEEQVQTTLSVT
Ig3
APLKAEVHPAVQKADVGRPATFNCTAAGHPVRSVSWYKDQTRLGSTSRLTLLASGHVLRIDSVLREDAGMYQCYLHNEAD
SAQASAELLLGD
Ig4 ? incomplete
VAPFLSSSFAEQTLSPGATLSLRCAAVGSPIPQVTWKLDGGPVPDLARFRVGDFVTSDSVV
Ig5
VSFVNVTEIRVEDGGEYACSASNVVGDVVHAARIDVHGP
Ig6
PTVRSMGNITVVAGTLLRIICPVSGYPIHGVGWFKGEQSYPCLRARFSMTHATLTVQNVQRASDEGEYACVARSGNLSAQ
GNTFVHVQ
Ig7
VPPVIDSQSLPDVLTANQGMNVKMLCSVVQGDPPISLRWFRGGNVVSRSASVSLQSLEDSSVLTLKGVVMRDSGNYTCVA
SNRAQAVNKSVTLVVN
Ig8 incomplete
AEPRNFDLVSSSYRVQILSNGSLVIQDTELGDGGYYLCEAHNGIGVGLSRVIALSVN
Ig9
VPPSFSTKFSSHNVKRGGQEAVLRCEAKGDPDLEITWEKDKHPMDLTTEKRYSLTEDTSRNRMSSSLTILLTERRDGALYS
CIARNPFGSDETNIQLLVQ
FNIII-3-1
EAPSAPAEVRISKVASRTLEISWSPSYN**GNNP**IRKYHVHFTNSTSSWDSTSSRLQLSVPGTETKATIHKLHPVTTYRIRV
TAENMPTYVTL
FNIII-3-2
CPLQPPRKDLHHGKVQ**GYYIGY**KEVEKEEAEFQYKNVEALDVTSGARLHQMSH
LTNLKRKTSYVVKVQAYNSEGAGPMSDDVRATTLEA
FNIII-3-5
DYVLHYQVKGGDWQQKALSTNSNKYTVEGLKCGSVYSLYMTATNSLGTAEPRDIIYARTKGADDPLSKCRGSVDNMGMAE
FCAM
KQRLQQQQLRHKEEEAYSKGTSFYASPARKPVPVSSDPRM

Is53 contig 645963

Ig 7.1
VALTVRIQPFVFPEKAVV**G**TKVSVMCTTVEEIPTVQFR**W**YKNGSPLVTSESNSRVRLRTFPDVSNLVIGPLEEGDSGNYT
CTGTTKSRSDSHTEVLSVL
Ig8.1
VPPKWIHEPQDANLREGQNLSVRCEAKGHPTPTVQWKLKGNRNVAMANDSRGGLLTISKATKDVAGTYVCTADNGLPDKL
SREIRINIFGENSP
Ig7.2
RLQPFSLPTDAIEGNKVTATCAPVTGGISSGIEFSWFKDGRKLVQDGRVRVRSFPDMSTLVVDTLKQEDSGNYTCVGKLR
NQKDSHTEVLRVL
Ig8.2
VPVKWLREPADVFLRETENATLSCEATGVPKPTVKWDKEETESKHSFYAGIALSAQSSSLMLFKATKGDAGNYRCTADNG
LRNRLTKRIRV
Ig7.3
RLQPFHFPTDAVEGKTVTVLCTTTTAITGVEYRWLKNNKRVTENSKIRLRTFPELSSLIVGPLEAFDSGNYTCQGLYNGK
KDSFSDTLNVL
NNNNNNN
Ig7.4
AGALHLQPFTFPSKVVEGTTVTVLCTTTSGIANVNFRWLKDGREIATSAKVKIIHHSLLSSLVIGPVNRGDSGNYTCVGN
IGEKLNSHSEVLSVL
Ig8.3
APPEWIVEPEDIKVHQGGNVTIACEAAGNPTPTVKWTQRYQLTTESLSKYLNLYVTNWRSLTGPAKETSASGRNLLTLAN
ASKSDAATYECRAVNGVPEDIYKRV

Ig7.5
ALVVKVQPFSFPNDLLEGSRVSVACSLRKVSSDARFKWLKDGKALDGNRYRRLSVRTEADFSMVTIDPARQEDSGNYTCT
VTSKGRSDSYTAALVV
NNNN
Ig7.6
VPPKIHPFAFSKVLVVGERSSVTCTTIAGDKPFKFIWLKDGSTLRQEGNVKIVSSSEFSMFNIERLSLENAGNYTCVVSN
AGGTVSYASTLEI
g8.4
APPTWKTEPRDMSVTAGQKVSITCDGNGHPQPSVRWTKEGTLLCGSDGGSSEYRTKTIELASASKPDQGSYTCEIANGIG
EAIRKTITILIg7.7
IPPFQFPKNLQVDQRISVICTISV**GD**T**P**IQFAWL**KDG**SALSNSSPNVRIVDSAEFSTLHIAPLTLNSA**GNYT**CSV**SN**K**A**G
YTSY**T**APLV**V**

Ig8.5
APPRWIKEPQDVTATAGSNVTMACSADGFPKPSVNWRKLESDSESPTPVIEPHLDHKGTSTIAIVSVGKLHQGRYSCLVS
NGIGLDLSKTVSLRI
Ig7.8
APKIQPFTFPTTLNAGERTATICVVTAGDKPLTFSWFKDGKTLETEDNVKITSNAEFSNLNFGSLTVKHSGNYTCSVKNN
VGSASFTAAFLAV 48898
NNNNNN
Ig8.6
AAPPEWKTEPRDLSVSAGQALLVECSATGYPLPKVTWKKDGPKNEQTLVASQDGSATLSVTESTKETEGRYFCEADNGVG
AALKTALFIKNNNNNN
Ig7.9
LNDFSEAPKIQPFTFNDKVRI**G**GRAVGSCIVVTAAAPLTFT**W**IKDGVQLRDKTGLSIQNNRLVSLLIIETADVFSH**G**N**Y**T
CRASNAMGTDAYTAELKVE
Ig8.7
APPTWSHEPQDVSAIVGTNVTVE**C**RATGSPIPQITWTKSKGKFLHQSWKRTASKNIHPDVLCITDGTSTRLLMHKDVLLI
QNIQDVDAGS**Y**TCKAENGVGPTLHKTVRV
Ig7.10
PKVLPFNFLKTLSEGQSALVTCTVSEGSKPVQLQWLKDGHEVRASSTVKIKRDETFVVLAIEPVQVEDSGNYTCVAKNKY
GYDRYTSLLEVH
Ig8.8
KYGWFVFVLLAPPKWLHEPSDVALTSGEAAMLHCKAAGHPTPSIKWSRSGTEGSSRVLENGTFIISKSAPEDTGQYSCQA
SNGIGNincomplete due to NN

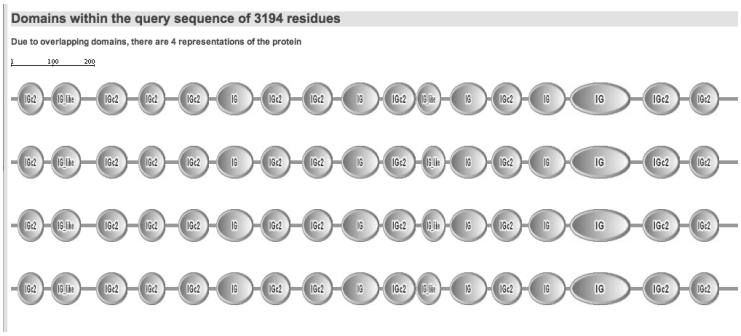


Figure S6 Hemocytes withdrawn from *S. maritima* and stained with Giemsa.

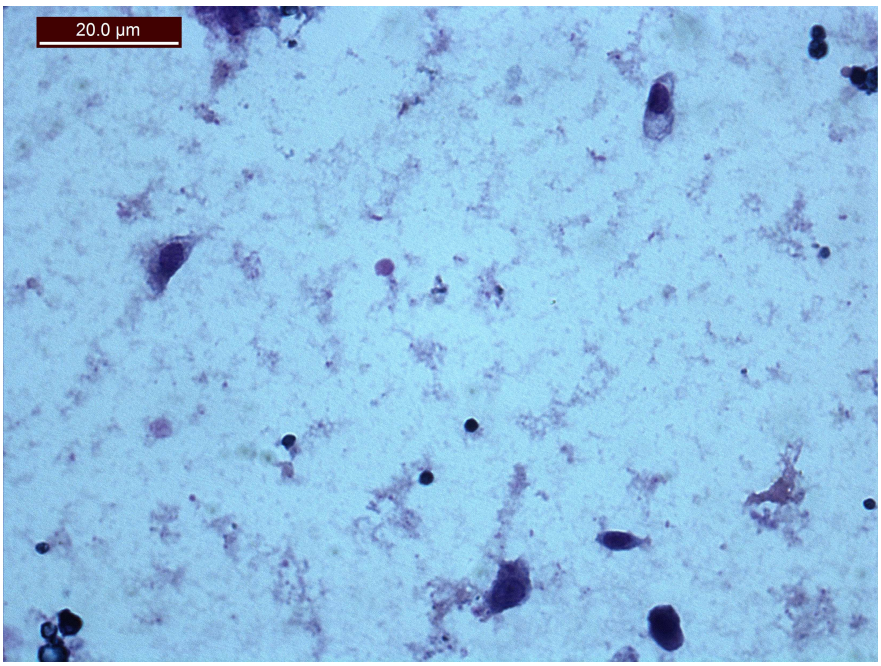


Figure S7 Alternative splicing of *Sm35* cytoplasmic tail from *Strigamia maritima*. The number on the right of *Sm35* refers to transcripts analyzed. In red are represented exons alternatively spliced; in total, three different cytoplasmic tails were found to be expressed.

```
Sm35_8934-4    GRIINGIVKSSSKFSSSSSTVKNYSVDSFGNGQRSTESIARRYPSISDKLVKEIVCSWNL
Sm35_8934-6    GRIINGIVKSSSKFSSSSSTVKNYSVDSFGNGQRSTESIARRYPSISDKLVK--------
Sm35_7383-2    GRIINGIVKSSSKFSSSSSTVKNYSVDSFGNGQRSTESIARRYPSISDKLVK--------

Sm35_8934-4    GLFVSKSQDCKMKYCTWEVCFFLILSRIMSRTKKAILLDSFIIHRDSPRSRSAPGSSDEI
Sm35_8934-6    ----------------------------------------------------GSSDEI
Sm35_7383-2    ----------------------------------------------------GSSDEI

Sm35_8934-4    TPYATTQLPNFHYGEMKTFGERKSGASPFSGGGGSDNEENLIQNTNTQKRVKKQSGEQIA
Sm35_8934-6    TPYATTQLPNFHYGEMKTFGERKSGASPFSGGGGSDNEENLIQNTNTQKRVKKQSGEQIA
Sm35_7383-2    TPYATTQLPNFHYGEMKTFGERKSGASPFSGGGGSDNEENLIQNTNTQKRVKK-------

Sm35_8934-4    RPKSDGAVVAAAYPRPEPDGKAAWATGQPERGFSSQTGFPVGQSRSAARLPDSSMTRANS
Sm35_8934-6    RPKSDGAVVAAAYPRPEPDGKAAWATGQPERGFSSQTGFPVGQSRSAARLPDSSMTRANS
Sm35_7383-2    ------------------------------------------------------------

Sm35_8934-4    GGPSPRQQASPGDTKWRIVQRNLGNISKAKVHGVGSSSGTQETTFIFPRTPDEVGVTPTM
Sm35_8934-6    GGPSPRQQASPGDTKWRIVQRNLGNISKAKVHGVGSSSGTQETTFIFPRTPDEVGVTPTM
Sm35_7383-2    ------------------------------------------------------------

Sm35_8934-4    MSSDPTERYDEPILPPSAFQNKGKTDQTQADPTEGSKLLKRSLVSCK
Sm35_8934-6    MSSDPTERYDEPILPPSAFQNKGKTDQTQADPTEGSKLLKRSLVSCK
Sm35_7383-2    -----------------------------------------------
```

Figure S8 Maximum Likelihood topology depicting the phylogenetic relationship between Ig7 coding exons (n=178) for different Dscam from different species. Bootstrap values only significantly (>60%) for the branches in red. Paralogous exons within species were collapsed for simplicity.

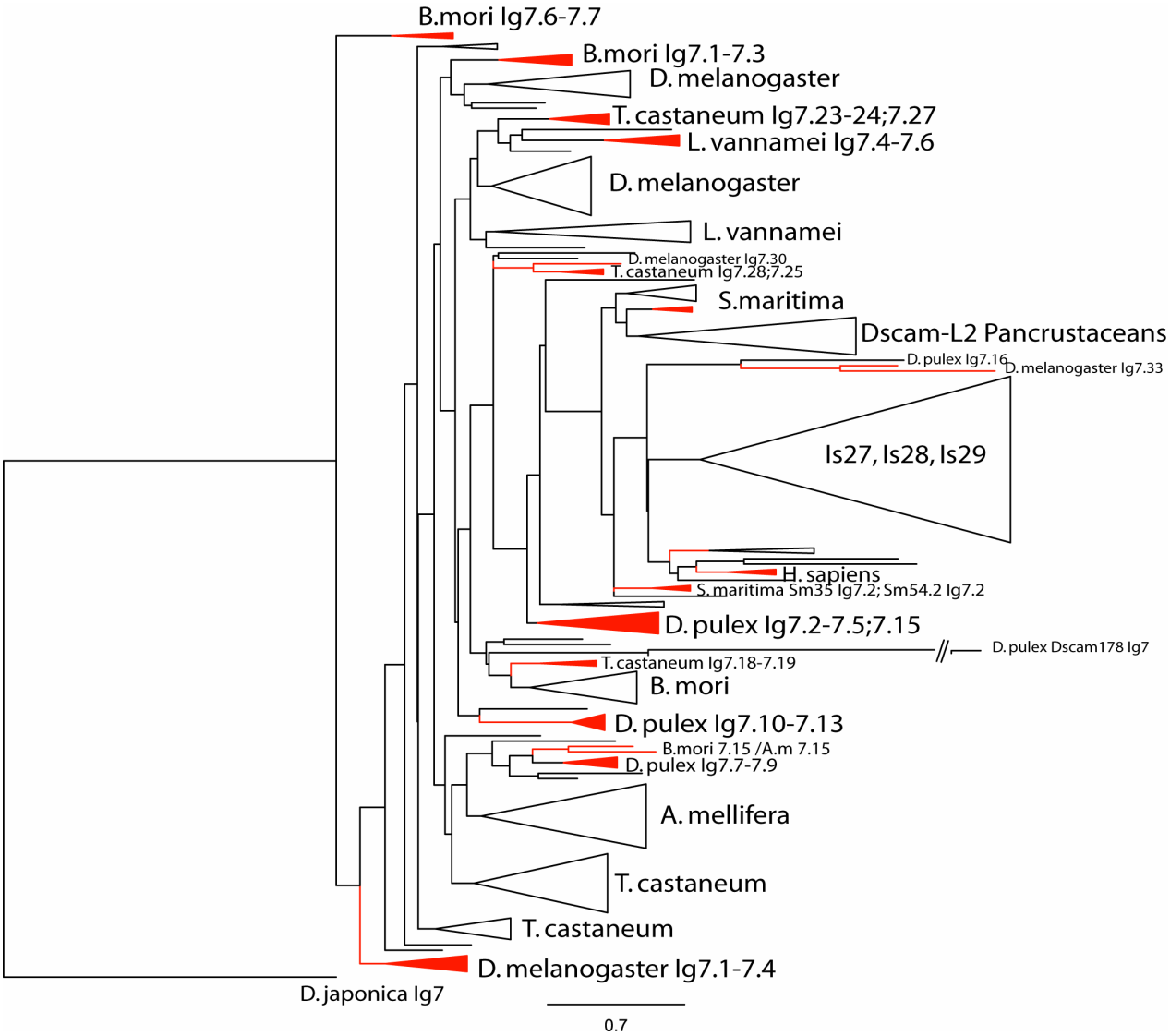


Figure S9 Representation of the amino acid conservation of exons coding for Ig7 of Dscam-hv of 6 pancrustacea species and of all other Dscam homologues in the remaining species (Table S1). Hallmark amino acid position of Ig7 domains are marked (*) and numbered.The size of the letter is proportional to the frequencies of each amino acid in each position. The colors represent the chemical properties of amino acids; polar (green), basic (blue), acidic (red) and hydrophobic (black). This figure was created with WebLogo (http://weblogo.berkeley.edu/logo.cgi).

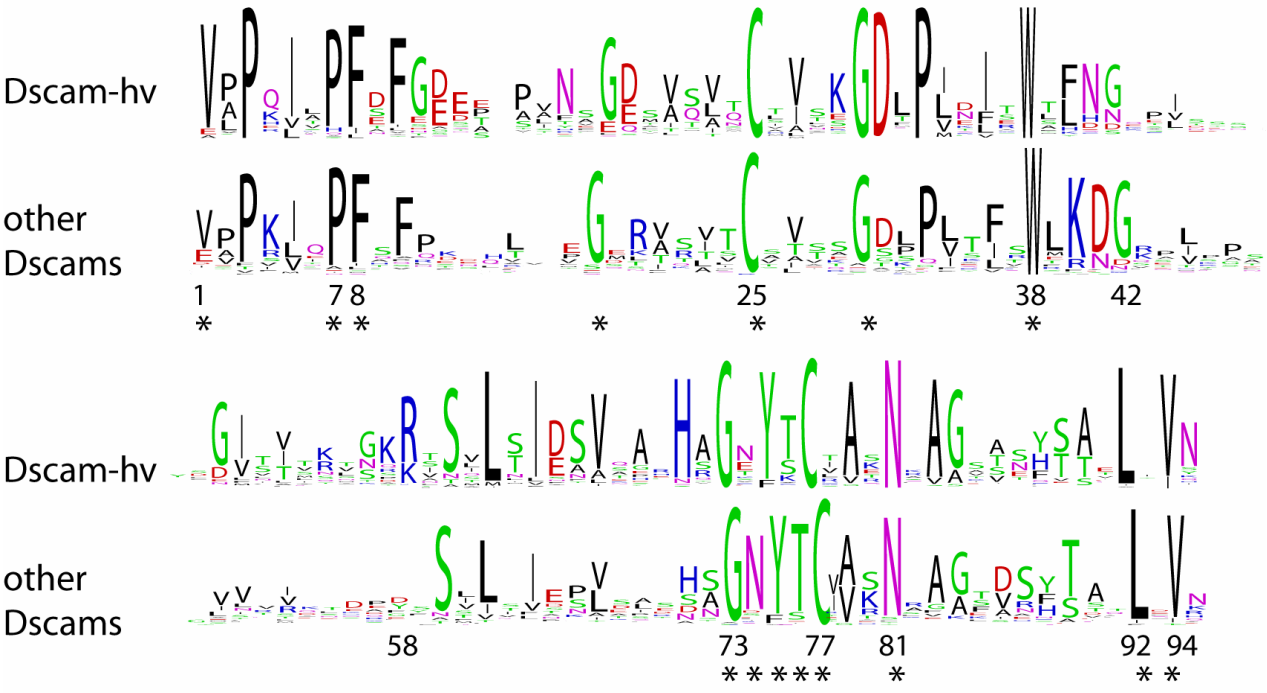


Table S3 Summary of the cytoplasmic tail motifs found in reconstructed Dscam homologues of several species.x stands for any amino acid; ( ) indicates motifs that are not canonical. [1] Internalization motifs; [2] n° of cysteines

Species/Dscam member	Length (n°aa)	TM Association possibility[2]	SH2 Binding sites	YxxL[1] YxxI	ITIM	ITAM	Other peculiarities
Sm54.1	165				(1)		
Sm 54.2	72	3 C	1				
Sm 24	204	3C	6				1 YxxQ (STAT3 phosphorylation)
Sm 34	151	1C	1		(1)		
Sm 52	89	1C	1				
Sm166	60	4C					
Sm54	84	1C			(1)		
Sm82	56	1C					
Sm29	186	3C	3	1			1 YxxQ
Sm32	211	1C	1				2 YxxG (endocytosis); Polyprolin
Sm17	213	1C	1				12 Ys; 2YxxF (trafficking)
Sm16	339	1C	1	1			
Sm14	60	1C	1				1 YxxG
Is3	2	1C					
Is4	148	2C	4				
Is8	117	3C	2 (2)				
Is10	218	1C	6 (1)		(1)		1YxxY (STAT3)
Is15	131	1C	3				
Is17	81						
Is23	219		3				1 YxxF
Is26	16	1C					
Is32	171	1C	2				
Is27 to Is28	443		4	1	1 (1)		1 YxxG
Is20	43	1C					
Is6	101		5				1 YxxG
Is22	274	1C		1			1 YxxG
Is25	160	2C		1	(1)		
DSCAM (human)	264 short 304 long	1C	4		1	1	1YxxF: polyprolin
Nematostella 1	458		5		1		2 YxxG; polyprolin
Nematostella 2	384		2	1	1		1 YxxQ
Nematostella 3	448		5				Polyprolin
Nematostella 4	371		13				
Sp (sea urchin)	354		3	1			1 YxxF

CHAPTER 5

OUTLOOK

Despite the fact that many functional aspects of the Dscam gene are still unknown, its role in the nervous system has been elucidated over the last decade in great detail; essential functions have been described, the role of isoform diversity in creating binding specificities is understood, the molecular structures underlying the specificity of binding have been discovered. These are few out of a much larger list of achievements made by several groups and different lines of work.

Contrastingly, much remains to be done to understand the function of Dscam in immunity. Several fundamental questions remain unknown and untested. For instance, how do the different isoforms act in the context of an immune function? Is the repertoire of certain isoforms amplified under infection? Is that due to up-regulation of the gene or does cell proliferation play a role? An important question that needs to be investigated is whether there is specific proliferation of hemocytes after infection. In this respect, some differences between crustaceans and insects might be expected based in what is known about hematopoiesis in representatives of both groups. In *Drosophila melanogaster*, all circulating adult hemocytes are of larval origin and a certain part of the larval produced hemoytes is stored and released under parasite challenge (Wood and Jacinto 2007). This aspect of *Drosophila* hematopoeisis invalidates to a certain extant models proposed for the action of Dscam as an immune receptor (Boehm 2007) given that clonal amplification of cells expressing a certain Dscam repertoire has not been demonstrated. The situation in crustaceans might be different, given that at least *de novo* proliferation of hemocytes in the hematopoeietic tissue of the cray fish *Pacifastacus leniusculus* and of the shrimp *Penaeus japonicus* has been suggested (Sequeira, Tavares, and AralaChaves 1996; Soderhall et al. 2003). However, there is still no convincing demonstration of specific hemocyte proliferation, i.e. production of hemocytes with properties enhanced by a certain elicitor.

The observations that there is no general up-regulation of the Dscam gene under infection, if they hold true, are also puzzling because that would imply that the total amount of expressed Dscam does not increase under infection and perhaps only qualitative changes on the repertoires of exons transcribed take place. Could an amplification of certain Dscam repertoires happen at the level of the soluble forms produced by hemocytes and/or by the hematopoietic organs, by maintaining Dscam expression constant and regulating splicing of the alternative exons? More experiments are needed to understand this fundamental aspect of the immunobiology of Dscam, namely testing whether specific proliferation of hemocytes can

occur, and investigating whether regulation of alterative splicing during an immune response takes place. The former could be done by comparing molecular markers of Dscam or other genes in new populations of proliferating hemocytes in control and challenged individuals. A large crustacean would be possibly the most suitable model for such experiments given that hemocyte proliferation seems to occur in these animals, and large amounts of hemolymph can be withdrawn. Among insects, bigger species and living longer than *Drosophila* or *Anopheles* such as the bumblebees, might give additional interesting insights.

The question of whether alternative splicing is regulated during an immune response could be approached by obtaining a robust representation of all Dscam transcripts expressed in animals under a parasite challenge compared with controls. High throughput sequencing techniques would allow analyzing several replicates which would strongly enhance the significance of the results. *Daphnia magna* would be an ideal model system for carrying out such experiments given that genetic and developmental differences between individuals and replicates can be nearly entirely controlled by replicating clonal individuals. The use of replicated clones could further help elucidating whether expression in brain and hemocytes of control and challenged animals is arbitrary (replicates would express different repertoires) or deterministic (replicates would express similar repertoires). If the expression of repertoires is arbitrary that would suggest that only Dscam diversity matters but not the nature of its diversity. Contrarily, if expression is deterministic it would be an indication that the exact amino acid composition of the variable regions is important. This would have profound implications in our understanding of the Dscam function in both the nervous and immune systems.

The present and other studies provided candidate exons and/or exon associations (Dong et al. 2006; Brites et al. 2010), whose binding affinities to different antigens could be tested by binding *in vitro* Dscam constructs with a certain exon composition to different parasites and pathogens. The strength of binding could be further assessed by blocking or modifying the Dscam epitopes supposedly involved in parasite recognition (Meijers et al. 2007), by using antibodies and by site-directed mutagenesis, respectively. Another aspect that needs more investigation is the function of the Dscam soluble isoforms. Despite the suggestive evidences that they might be expressed in crustaceans besides insects (Chou et al. 2009) and in *Ixodes scapularis* and *Strigamia maritima*, there is still no confirmation for that at the protein level. It also remains to be shown whether Dscam soluble forms in the hemolymph bind *in vivo* to the hemocyte surface Dscam receptors and to antigens.

There is mounting evidence that at least some groups of arthropods exhibit immune phenomena such as specific memory thought to be unique to vertebrates. Such phenomena could

be explained by immune priming, a persistent state of an immune function, specific or not, after a first encounter with an antigen (Kurtz and Franz 2003; Sadd and Schmid-Hempel 2006; Roth and Kurtz 2009) In some cases, the responses found revealed a high degree of specificity, implying the ability for distinguishing between gram-positive and gram-negative bacteria or even between strains of a same parasite (Roth and Kurtz 2009). A comprehensive view of the immune functions underlying such responses is lacking but there are evidences in different taxa for an involvement of phagocytosis (Pham et al. 2007; Roth and Kurtz 2009). Therefore Dscam, mainly due to its extreme ability to generate diversity and its reported strong effects on phagocytosis, has been put forward as an exciting candidate for mediating specific immune responses in Arthropods (Kurtz and Armitage 2006). Nevertheless we are still far from understanding how that could happen. One hypothesis is that the soluble forms of Dscam, after binding to foreign epitopes, interact with the Dscam membrane bound isoforms of hemocytes via homophilic binding (Meijers et al. 2007). This could trigger the formation of multiprotein assemblies that lead to cellular uptake reactions such as phagocytosis. The amplification of the response could be at the level of these multiprotein assemblies which could activate cellular uptake in other hemocytes where Dscam homophilic binding between soluble and membrane forms would not occur. The interaction of multiprotein assemblies with other cell adhesion molecules such as hemolin has been put forward as an important component of arthropod cellular immune reactions (Schmidt et al. 2010). Multiproteins assemblies have been furthermore suggested, to be a possible mean of generating specific immune responses (Schulenburg, Boehnisch, and Michiels 2007). Such a scenario could explain how a certain level of specificity could happen in the absence of clonal expansion of Dscam isoforms elicited by a pathogen challenge.

The genetic diversification of the Dscam gene is exploited by the nervous system and perhaps by the immune system. Immunoglobulin domains are part of many cell adhesion molecules of the nervous and immune systems in vertebrates and invertebrates (Brummendorf and Lemmon 2001). But a common usage by both systems of a high diversity of receptors encoded by the same locus is a remarkable feature of Dscam (Du Pasquier 2005). How did this duality evolved? Given the conserved role of Dscam in the nervous system, perhaps the most parsimonious hypothesis is that diversification created by duplication and alternative splicing was initially exploited by the nervous system. The involvement in immunity might have appeared later, profiting from expression of Dscam diversity by hemocytes. That could have been (could be) advantageous in the context of cell migration during embryonic development, and hemocyte circulation in the hemolymph of adults. But given that in the ancestors of

pancrustaceans a non variable Dscam was likely already used by the nervous system, another attractive hypothesis is that hemocytes profited initially from isoform diversity and that was followed by the involvement in the nervous system.

The study of Dscam in other basal arthropod organisms, both by investigating Dscam expression in different tissues and by inferring functional constraints from molecular evolution patterns between different Dscam family members, will certainly bring interesting insights into this issues.

Other aspects of Dscam to be further studied are summarized in Table 1. Dissecting the function and evolution of this gene will be a challenging endeavor. However, that might be rewarded by improving considerably our understanding of the nervous and immune systems of arthropods, and our understanding of how evolution has built this extremely complex solution to serve these two systems.

Dscam feature	To be tested
Signalling:	Signal transduction pathways
	Role of ITIM and ITAM
	Cytoskeleton connections
	Role of PDZ motifs
Transmembrane domains:	Role of cyteines
	Multiprotein associations
Receptor:	Isoform specificity
	Surface expression
	Soluble forms
	Cellular localization
Role in immunity:	Effect of knockout
	Binding to antigens and parasites
	Kinetics of expression
	Alternative splicing
	Fat body *vs* hemocytes
	Function in other animal models
	Hemocyte circulation
Evolution:	Dscam in other arthropods
	Dscam in pre-bilateria members
	Expression in different phyla
	Relationship to other CAMs which form a horse-shoe structure

Table 1 - Aspects of Dscam to be further investigated, suggested from this and other studies

REFERENCES

Boehm, T. 2007. Two in one: dual function of an invertebrate antigen receptor. Nat Immunol **8**:1031-1033.

Brites, D., F. Encinas-Viso, D. Ebert, L. Du Pasquier, and C. R. Haag. 2010. Signatures of selection on duplicated alternatively spliced exons of the Dscam gene in *Daphnia* and *Drosophila*. *in preparation*.

Brummendorf, T., and V. Lemmon. 2001. Immunoglobulin superfamily receptors: cis-interactions, intracellular adapters and alternative splicing regulate adhesion. Current Opinion in Cell Biology **13**:611-618.

Chou, P. H., H. S. Chang, I. T. Chen, H. Y. Lin, Y. M. Chen, H. L. Yang, and K. C. H. C. Wang. 2009. The putative invertebrate adaptive immune protein Litopenaeus vannamei Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. Developmental and Comparative Immunology **33**:1258-1267.

Dong, Y., H. E. Taylor, and G. Dimopoulos. 2006. AgDdscam, a Hypervariable Immunoglobulin Domain-Containing Receptor of the *Anopheles gambiae* Innate Immune System. PLoS Biol **4**:e229-.

Kurtz, J., and S. A. Armitage. 2006. Alternative adaptive immunity in invertebrates. Trends Immunol **27**:493-496.

Kurtz, J., and K. Franz. 2003. Evidence for memory in invertebrate immunity Nature **425**:37-38.

Meijers, R., R. Puettmann-Holgado, G. Skiniotis, J.-h. Liu, T. Walz, J.-h. Wang, and D. Schmucker. 2007. Structural basis of Dscam isoform specificity. Nature **449**:487-491.

Pasquier, L. D. 2005. Insects Diversify One Molecule to Serve Two Systems. Science **309**:1826-1827.

Pham, L. N., M. S. Dionne, M. Shirasu-Hiza, and D. S. Schneider. 2007. A specific primed immune response in Drosophila is dependent on phagocytes. PLoS Pathog **3**:e26.

Roth, O., and J. Kurtz. 2009. Phagocytosis mediates specificity in the immune defence of an invertebrate, the woodlouse Porcellio scaber (Crustacea: Isopoda). Dev Comp Immunol **33**:1151-1155.

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

Schmidt, O., K. Soderhall, U. Theopold, and I. Faye. 2010. Role of Adhesion in Arthropod Immune Recognition. Annual Review of Entomology **55**:485-504.

Schulenburg, H., C. Boehnisch, and N. K. Michiels. 2007. How do invertebrates generate a highly specific innate immune response? Molecular Immunology **44**:3338-3344.

Sequeira, T., D. Tavares, and M. AralaChaves. 1996. Evidence for circulating hemocyte proliferation in the shrimp Penaeus japonicus. Developmental and Comparative Immunology **20**:97-104.

Soderhall, I., E. Bangyeekhun, S. Mayo, and K. Soderhall. 2003. Hemocyte production and maturation in an invertebrate animal; proliferation and gene expression in hematopoietic stem cells of Pacifastacus leniusculus. Developmental and Comparative Immunology **27**:661-672.

Wood, W., and A. Jacinto. 2007. Drosophila melanogaster embryonic haemocytes: masters of multitasking. Nature Reviews Molecular Cell Biology **8**:542-551.

AKNOWLEDGEMENTS

I feel the luckiest and happiest of all students to have worked with Louis Du Pasquier. This is certainly one aspect of my PhD that I will never forget. Thank you Louis for sharing your knowledge with me, for your guidance, for your modesty, for your patience and for many other things that I cannot name easily.

Nothing would have been possible without the support of Dieter Ebert. I learned a lot with him, he always guided me when I needed and gave me at the same time all the freedom I wanted. I admire him for that and I feel very grateful.

I would like to thank my husband Philipp for encouraging me all the time and being always there when I needed him. My daughter Clara was born in the beginning of 2009, since then I sleep less but smile more. *Obrigada* Clarinha.

My parents in law helped me enormously by taking care of Clara very often so that I could progress, I am very thankful to them for that.

I would also like to thank all my colleagues which help me in one way or the other with ideas, statistical analysis, *Daphnias* for experiments, lab support, etc, and not least, for being nice people whom is great to meet everyday; Jürgen Hottinger, Urs Stifel, Brigitte Aeschbach, Lukas Zimmermann, Dita Vizoso, Lucas Shärer, "the girls office"; Flore Mas, Frida Ben-Ami, Kyono Sekii, Karen Haag and Nicolas Boileau (it was not the girls office at that time!); Thomas Zumbrunn, David Duneau, Pepijn, Thomas Frabbro, Sandra Lass, Olivia Roth, Francisco Encinas-Viso, Harris, Adrian Baummeyer, Isabelle Colson, Florian Altermatt and probably others….

I would also like to thank other people with whom I collaborated and learned a lot, Seanna McTaggart, Dietmar Schmucker and Christoph Hagg. Dietmar Schmucker hosted me during one month in his lab and that was a great experience. I would like to thank him for that as well.

THANK you to all my friends and my family for being just what they are.

I thank "Quasilusos" for being the craziest and funniest theater group I could find in these latitudes and for being very dear people.

Finally, I would like to thank Hinrich Schulenburg for having accepted to referee my PhD.

I would like to thank the *Fundação para a Ciência e Tecnologia* for funding most of my PhD and my research. I was also funded by the *Roche Science Foundation* and by the *Reise Fonds* of the University of Basel. I am very thankful for their support.

Curriculum vitae

Name	Daniela Alexandra da Silva Henriques Brites
Address	Jacobistrasse 10, 79104 Freiburg (Germany)
Phone	++497612088615
Email	danielabrites@yahoo.com
Nationality	Portuguese
Date of birth	25.06.1977

EDUCATION

2005 - 2010	PhD at the University of Basel. Supervision: Prof. Dieter Ebert and Prof. Louis Du Pasquier. Thesis: Evolution and expression of the highly variable cell adhesion molecule Dscam in the crustacean *Daphnia* and other arthropods.
2000 - 2001	Diploma thesis, Center of Environmental sciences, Madrid, Spain Supervision: Dr. Fernando Valladares. Thesis: Symmetry and mathematics of plant foliage: curiosity or function? The influence of phyllotaxis in light harvesting of 12 Mediterranean woody species assessed with a 3-D computer model.
1996 - 2001	Graduate education in Biology at the Science College of the University of Lisbon, Portugal.

WORKING EXPERIENCE

2011-2012	Postdoc researcher at the Swiss Tropical and Public Health Institute, Basel
2007	Quantification of RNA expression by quantitative PCR. Dana Farber Cancer Institute from University of Harvard, USA. P.I. Dietmar Schmucker. (1 month).
2004 – 2005	Research assistant on experimental evolution and molecular biology. Autonoma University of Barcelona, Spain. P.I: Dr. Mauro Santos. (6 months)
2003 - 2004	Research assistant on experimental evolution and the genetic basis of adaptation. Gulbenkian Institute of Science, Oeiras, Portugal. P.I.: Dr. Henrique Teotonio.
2001-2003	Field and research assistant on plant ecology and ecophysiology. Center for Environmental Sciences, Superior Council of Scientific Research (CSIC). Madrid, Spain. P. I. : Dr. Fernando Valladares

TEACHING

2008	Teaching assistant in the block course of Zoology and Evolution , University of Basel
2007	Teaching assistant in evolutionary genetics, University of Basel
2006 – 2007	Teaching assistant, practical course on Zoology and Evolution

GRANTS

2011	Post-doctoral Marie Heim-Vögtlin fellowship (Swiss National Foundation)
2008	Roche Foundation fellowship
2005-2008	PhD fellowship and research grant from the Portuguese Science Foundation
2005-2008, 2010	Travel funds from the University of Basel
2000-2001	Erasmus fellowship funded by the European Union

CONTRIBUTED TALKS AND POSTERS

2012	Research seminar, Department of Zoology, University of Cambridge, UK (invited talk).
2011	Conference Jacques Monod, Coevolutionary arms race between parasite virulence and host immune defence: challenges from state of the art research. Roscoff, France (poster).
2010	Evolutionary and ecological genomics of adaptation, University of Fribourg, Switzerland (poster)
2010	Research seminar, National Institute for Medical Research, London, UK (invited talk).
2009	EMBO course on Molecular Tools on Development and Evolution, Kristineberg marine station, Sweden (poster).
2008	Annual Meeting of the Society for Molecular Biology and Evolution. Barcelona, Spain (poster).
2007	11th Congress of the European Society of Evolutionary Biology. Uppsala, Sweden (talk).
2007	European Foundation Conference, The impact of environment in innate immunity. Obergurgl, Austria (poster).
2007	Conference Jacques Monod, Evolutionary genetics of host-parasite relationships. Roscoff, France (poster).
2007	III Portuguese Meeting of Evolutionary Biology. Gulbenkian Institute of Science, Oeiras, Portugal (talk).
2006	II Portuguese Meeting of Evolutionary Biology. CBIO, Vairão, Portugal (talk).
2006	Interaction Seminar, ETH Zürich, Switzerland (invited talk).

2005 11th Meeting of PhD students in Evolutionary Biology. Bourdeaux, France (talk).

COURSES AND WORKSHOPS

2011 Perl programming in biomedical research- Swiss Institute for Bioinformatics, Lausanne, Switzerland

2011 Unix programming - Swiss Institute for Bioinformatics, Lausanne, Switzerland

2010 Molecular evolution workshop, Woods Hole, USA.

2009 EMBO course on Molecular Tools on Development and Evolution, Kristineberg marine station, Sweden.

2008 Summer Computational Phyloinformatics Course – modules R and HYPHY, NESCENT, Durham, USA.

2008 Metchnikoff's Legacy in 2008, Institute Pasteur, Paris, France.

2007 Phylogeny and Evolution using Bioinformatics - European Molecular Biology network course. Lausanne, Switzerland.

2005 Autumn School in Evolutionary Medicine - Humbolt-University of Berlin. Berlin, Germany

2005 Guarda Workshop in Evolutionary Biology - University of Basel. Guarda, Switzerland.

PUBLICATIONS

Brites D., Gagneux, S. 2011. Old and new selective pressures on Mycobacterium tuberculosis. Infection, genetics and evolution. *In press*

Brites D., F. Encinas- Viso, D. Ebert, L. Du Pasquier and C. Hagg. Population genetics of duplicated alternatively spliced exons of the Dscam gene in *Daphnia* and *Drosophila*. 2011. PLoS ONE 6 (12): e27947. doi:10.1371/journal.pone.0027947

Brites D., S. McTaggart, K. Morris, J. Anderson, K. Thomas, I. Colson, T. Fabbro, Tom J. Little, D. Ebert and L. Du Pasquier. 2008 The Dscam Homologue of the Crustacean Daphnia is Diversified by Alternative Splicing Like in Insects. Molecular Biology and Evolution 25 (7):1429-1439.

Santos M., D. Brites & H. Laayouni. 2006 Thermal evolution of pre-adult life history traits, geometric size and shape, and developmental stability in *Drosophila subobscura*. Journal of Evolutionary Biology 19 (6): 2006-2021.

Brites D., F. Valladares 2005 Implications of opposite phyllotaxis for light interception efficiency of Mediterranean woody plants. Trees 19: 671-679.

Valladares F., D. Brites 2004 Leaf Phillotaxis: does it really affect light capture? Plant Ecology 174: 11-17.