The immunoglobulin superfamily in *Caenorhabditis elegans* and *Drosophila melanogaster*

A recent report by Vogel et al. describes a bioinformatic analysis of immunoglobulin superfamily (IgSF) members in *Caenorhabditis elegans* and *Drosophila melanogaster* (Vogel et al., 2003). We have previously published reports presenting genome-sequence-driven analyses of worm and fly IgSF members (Hutter et al., 2000; Hynes and Zhao, 2000; Aurelio et al., 2002). In Vogel et al. (Vogel et al., 2003), these papers are either not cited (Hynes and Zhao, 2000; Aurelio et al., 2002) or their content is essentially ignored (Hutter et al., 2000), although they cover much of the same ground as the Vogel et al. paper. Furthermore, the Vogel et al. paper contains errors and misclassifications of IgSF family members. Given the high degree of interest in this superfamily, we wish to correct the errors and clarify any misconceptions caused by the conflation of structural features with functional characteristics by Vogel et al.

Errors and misinterpretations in the data set

Although in general, initial genome-wide analyses of protein families are rarely free of errors, we think that such errors should not be taken lightly in the context of a refinement of previously published analyses. We list below the errors that we noted.

(1) UNC-73 is shown in figure 3 of the paper as a secreted protein. This is incorrect. It is well established in both worms and flies (where the protein is called Trio) that this protein is an intracellular signal transduction molecule with nucleotide exchange factor activity (e.g. Bateman and Van Vactor, 2001; Kubiseski et al., 2003; Newsome et al., 2000; Steven et al., 1998).

(2) UNC-89 is shown in figure 3 as being an extracellular matrix protein. However, it is a well-documented, intracellular muscle protein (e.g. Flaherty et al., 2002; Lin et al., 2003; Mackinnon et al., 2002).

(3) The F59F3.1, F59F3.5, T17A3.1 and T17A3.8 proteins have been published and are called VER proteins (Popovici et al., 2002; Popovici et al., 1999), which is not cited by the authors. Moreover, in figure 2, three of the four proteins are omitted.

(4) The T17A3.10 protein is incorrectly listed in the 'Cell surface – kinases and phosphatases' section of table 3. Although its extracellular domain is similar to VER receptor tyrosine kinases, T17A3.10 has neither a kinase- nor a phosphatase domain.

(5) Oig proteins, secreted 1-Ig domain proteins (Aurelio et al., 2002), are not shown in figure 3. They are listed in table 2, but without appropriate citation of the prior annotation.

(6) Beat Ia should be set apart from other Beat proteins in figure 3 as it has an additional domain (a Cysteine knot domain) (Pipes et al., 2001), not shown by the authors.

(7) The authors use unpublished information to state that K07E12.1 corresponds to DIG-1 (table 2), but they fail to acknowledge their source of information. Moreover, K07E12.1/DIG-1 protein has a plethora of domains characteristic of extracellular proteins, such as Sushi, EGF and vWF domains, and thus it should be classified as an extracellular protein, not as a protein of unknown cellular location.

(8) The authors are not consistent in their placement of molecules into distinct classes. For example, they define 'Cell surface proteins I' as transmembrane or membrane attached proteins (see p. 6320), yet, in table 2, list the secreted ZIG proteins ZIG-2 to ZIG-8 in the 'Cell surface proteins I' category. By contrast, in figure 3, the same ZIG proteins are shown as secreted.

(9) In figure 3 the structure for perlecan (UNC-52) is incorrect. As previously published, there are 17 Ig domains, two spaced near the amino end and then a cluster of 15 (reviewed by Rogalski et al., 2001). There should also be laminin G repeats, which are not shown in figure 3.

(10) The *C. elegans* Semaphorin 2a gene in table 3 (mab-20/Y71G12B.20) should be annotated as an experimentally characterized sequence (Roy et al., 2000).

(11) The authors claim to have identified 19 new Ig-proteins in *C. elegans*, as compared with their own previous analysis. Five of those (Y54G2A.25, C09E7.3, T19D12.7, F28E10.2 and T17A3.10) had been identified earlier by Hutter et al. or by Aurelio et al., and thus cannot be termed 'new' proteins (Hutter et al., 2000; Aurelio et al., 2002).

Classification of IgSF proteins

IgSF proteins are classified in this paper according to their domain organization. Although this is a useful classification from a structural point of view, it has only limited implications for the functions of the proteins. Treating these structural classes as being equivalent to functional classes is incorrect as members from each class have been shown to have overlapping functions. For example, most members of the 'Cell Surface I protein' class, classified as 'cell adhesion proteins' by Vogel et al., can clearly serve as signaling molecules (e.g. L1, NCAM, Robo and DSCAM) (reviewed by Rougon and Hobert, 2003). To illustrate one example, the Robo IgSF protein is classified as a cell adhesion protein by Vogel et al., yet it has clearly been demonstrated to be a signaling molecule acting through the recruitment of intracellular signal transducing molecules, such as kinases and nucleotide exchange factors (reviewed by Araujo and Tear, 2003; Dickson, 2002; Korey and Van Vactor, 2000; Patel and Van Vactor, 2002; Rougon and Hobert, 2003). Moreover, many proteins in Class I are not sufficiently well characterized functionally to support their classification as 'cell adhesion proteins'. Also, the vast majority of Class III molecules are not characterized functionally and may well have structural/adhesive roles, rather than signaling roles, as the authors imply. Consequently, conclusions made by the authors about the meaning of the expansion of 'functional' classes in *Drosophila* (see p. 6326, 'Proteins common and specific to *Drosophila* and *C. elegans*') are not justified. The lack of correct assignment of individual IgSF proteins also calls into question the claim of the authors that the particular nature of proteins of the *Drosophila* IgSF repertoire (see p. 6327) "must be one of the contributing factors responsible for, for example, the formation of a more complex cellular structure in *Drosophila*". Perhaps the most impressive case of expansion of the IgSF repertoire in *Drosophila*, the thousands of

alternatively spliced isoforms of the IgSF protein DSCAM (Schmucker et al., 2000), is unfortunately not mentioned by Vogel et al.

References

Araujo, S. J. and Tear, G. (2003). Axon guidance mechanisms and molecules: lessons from invertebrates. *Nat. Rev. Neurosci.* **4**, 910-922.

Aurelio, O., Hall, D. H. and Hobert, O. (2002). Immunoglobulin-domain proteins required for maintenance of ventral nerve cord organization. *Science* **295**, 686-690.

Bateman, J. and Van Vactor, D. (2001). The Trio family of guanine-nucleotide-exchange factors: regulators of axon guidance. *J. Cell Sci.* **114**, 1973-1980.

Dickson, B. J. (2002). Molecular mechanisms of axon guidance. *Science* **298**, 1959-1964.

Flaherty, D. B., Gernert, K. M., Shmeleva, N., Tang, X., Mercer, K. B., Borodovsky, M. and Benian, G. M. (2002). Titins in *C. elegans* with unusual features: coiled-coil domains, novel regulation of kinase activity and two new possible elastic regions. *J. Mol. Biol.* **323**, 533-549.

Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C., Mastwal, S., Zhu, X., Scheel, J. et al. (2000). Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989-994.

Hynes, R. O. and Zhao, Q. (2000). The evolution of cell adhesion. *J. Cell Biol.* **150**, F89-F96.

Korey, C. A. and Van Vactor, D. (2000). From the growth cone surface to the cytoskeleton: one journey, many paths. *J. Neurobiol.* **44**, 184-193.

Kubiseski, T. J., Culotti, J. and Pawson, T. (2003). Functional analysis of the Caenorhabditis elegans UNC-73B PH domain demonstrates a role in activation of the Rac GTPase in vitro and axon guidance in vivo. *Mol. Cell. Biol.* **23**, 6823-6835.

Lin, X., Qadota, H., Moerman, D. G. and Williams, B. D. (2003). C. elegans PAT-6/actopaxin plays a critical role in the assembly of integrin adhesion complexes in vivo. *Curr. Biol.* **13**, 922-932.

Mackinnon, A. C., Qadota, H., Norman, K. R., Moerman, D. G. and Williams, B. D. (2002). C. elegans PAT-4/ILK functions as an adaptor protein within integrin adhesion complexes. *Curr. Biol.* **12**, 787-797.

Newsome, T. P., Schmidt, S., Dietzl, G., Keleman, K., Asling, B., Debant, A. and Dickson, B. J. (2000). Trio combines with dock to regulate Pak activity during photoreceptor axon pathfinding in Drosophila. *Cell* **101**, 283-294.

Patel, B. N. and Van Vactor, D. L. (2002). Axon guidance: the cytoplasmic tail. *Curr. Opin. Cell Biol.* **14**, 221-229.

Pipes, G. C., Lin, Q., Riley, S. E. and Goodman, C. S. (2001). The Beat generation: a multigene family encoding IgSF proteins related to the Beat axon guidance molecule in Drosophila. *Development* **128**, 4545-4552.

Popovici, C., Roubin, R., Coulier, F., Pontarotti, P. and Birnbaum, D. (1999). The family of *Caenorhabditis elegans* tyrosine kinase receptors: similarities and differences with mammalian receptors. *Genome Res.* **9**, 1026-1039.

Popovici, C., Isnardon, D., Birnbaum, D. and Roubin, R. (2002). Caenorhabditis elegans receptors related to mammalian vascular endothelial growth factor receptors are expressed in neural cells. *Neurosci. Lett.* **329**, 116-120.

Rogalski, T. M., Mullen, G. P., Bush, J. A., Gilchrist, E. J. and Moerman, D. G. (2001). UNC-52/perlecan isoform diversity and function in Caenorhabditis elegans. *Biochem. Soc. Trans.* **29**, 171-176.

Rougon, G. and Hobert, O. (2003). New insights into the diversity and function of neuronal immunoglobulin superfamily molecules. *Annu. Rev. Neurosci.* **26**, 202-238.

Roy, P. J., Zheng, H., Warren, C. E. and Culotti, J. G. (2000). mab-20 encodes Semaphorin-2a and is required to prevent ectopic cell contacts during epidermal morphogenesis in Caenorhabditis elegans. *Development* **127**, 755-767.

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

Steven, R., Kubiseski, T. J., Zheng, H., Kulkarni, S., Mancillas, J., Ruiz Morales, A., Hogue, C. W., Pawson, T. and Culotti, J. (1998). UNC-73 activates the Rac GTPase and is required for cell and growth cone migrations in C. elegans. *Cell* **92**, 785-795.

Teichmann, S. A. and Chothia, C. (2000). Immunoglobulin superfamily proteins in *Caenorhabditis elegans*. *J. Mol. Biol.* **296**, 1367-1383.

Vogel, C., Teichmann, S. A. and Chothia, C. (2003). The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. *Development* **130**, 6317-6328.

Oliver Hobert[1,*], Harald Hutter[2] and Richard O. Hynes[3]

[1]Columbia University, Center for Neurobiology and Behavior, New York, NY, USA
[2]Max-Planck-Institut für medizinische Forschung, Heidelberg, Germany
[3]Howard Hughes Medical Institute, Center for Cancer Research, MIT, Cambridge, MA, USA
*Author for correspondence (e-mail: or38@columbia.edu)

Response

Looking at the bigger picture

Hobert et al. (Hobert et al., 2004) have made a number of criticisms on our paper (Vogel et al., 2003). In the following paragraphs we give our replies to these criticisms. In a number of cases, the comments do provide useful corrections to the paper. Nevertheless, the major conclusions of our paper are not affected by these corrections and they remain both novel and valid.

Previous work on the immunoglobulin superfamily repertoire

Prior to our work, Hynes and Zhao (Hynes and Zhao, 2000) stated that the number of IgSF proteins in *Drosophila* was about 150, and for about 130 of these they listed some or all of the domains by which they are formed. This information should have been cited in our paper and we regret not having

done so. Their results for *C. elegans* are similar to those we published (Teichmann and Chothia, 2000) prior to their paper. The work by Hutter et al. (Hutter et al., 2000) is acknowledged as a whole in our paper, but we are not able to make more detailed comparisons with our current work because their web database is inaccessible at present. The paper by Aurelio et al. (Aurelio et al., 2002) is discussed below.

Experimental characterisation of IgSF proteins

The most useful part of the correspondence by Hobert et al. (Hobert et al., 2004) is that which draws our attention to experimental work of which we were unaware. These correct the classification of two IgSF proteins: UNC-73 is an intracellular signalling molecule (Kubiseski et al., 2003) and

not a secreted protein; and UNC-89 is a muscle protein and not an extracellular matrix protein (Flaherty et al., 2002).

In addition, there are experimental papers on *C. elegans* IgSF proteins that we should have cited:

(1) Popovici et al. (Popovici et al., 2002) noted the homology of F59F3.1, F59F3.5, T17A3.1 and T17A3.8, and gave them the name VER proteins. This homology was also described, independently, by Teichmann and Chothia (Teichmann and Chothia, 2000).

(2) Aurelio et al. (Aurelio et al., 2002) characterised the expression of C09E7.3, Y38F1A.9 and Y50E8A.3, and gave the three proteins the names Oig-1, Oig-2 and Oig-3. [They also described the number of Ig and FnIII domains in 24 *C. elegans* proteins. Of these, four were claimed to be not present in the IgSF proteins listed by Hutter et al. (Hutter et al., 2000) and Teichmann and Chothia (Teichmann and Chothia, 2000). In fact, only two, C09E7.3 and Y42H9B.2, were new: the other two, the domain structure of Y38F1A.9 and Y50E8A.3, are described in figure 6 of Teichmann and Chothia (Teichmann and Chothia, 2000). The other sequence mentioned by Hobert et al. (Hobert et al., 2004), Y54G2A.25, is a revised version of Y94H6A_148.d that was used by both Hutter et al. (Hutter et al., 2000) and Teichmann and Chothia (Teichmann and Chothia, 2000).]

(3) The *C. elegans* Semaphorin-2a is an experimentally characterised sequence (Roy et al., 2000).

Hobert et al. (Hobert et al., 2004) correctly note that two Ig domains are missing from the Perlecan structure in figure 1 of our paper (Vogel et al., 2003). They also point out that whilst Zig-2 to Zig-8 are correctly described as secreted proteins in figure 3 of our paper, they are carelessly placed with Zig-1 in the cell surface category in table 2.

Classification of IgSF proteins

The classification of the IgSF proteins in our paper is based on their structural features, their subcellular location and sequence similarities. We give some rough descriptions of the more common functions of the proteins in the different classes. Hobert et al. strongly object to this (Hobert et al., 2004). They claim that we imply that *all* proteins in Class I are cell adhesion molecules. We actually say that the experimentally characterised proteins in this class are "*mainly* cell adhesion molecules". We and most readers are well aware of the multiple roles of, for example, Roundabout. Similarly, Hobert et al. (Hobert et al., 2004) claim that we imply that all Class III proteins are signalling molecules, whereas we state that "*those characterised so far* are signalling molecules".

Only a wilful literalist would take rough descriptions of the more common known functions to be precise descriptions of the functions for all the proteins in a class. Proteins with similar domain structures and related sequences do tend to have related functions (e.g. Hegyi and Gerstein, 2001). But, as we say in the paper (p. 6327), any type of function suggested for new proteins by their structural and sequence similarties to characterised proteins will need to be refined or corrected by experiments.

Conclusions

Many of the criticisms above are concerned with work by others that should have been cited. The criticisms of the results are in some cases correct but their overall effect is small. The more serious criticisms require that two proteins, UNC-73 and UNC-89, are placed in different classes, and that the secreted proteins Zig-2 to Zig-8 are placed in the correct part of table 3.

Because of the improvements in predictions of protein sequences made by the curators of the genome sequences, and because of improvements in sequence comparison procedures (Karplus et al., 1998; Gough et al., 2001; Madera and Gough, 2002), our descriptions of the IgSF proteins in *Drosophila* and *C. elegans* go beyond those published previously. The matches made by the sequence comparison programs are accompanied by a score that is an estimate of the match being in error. We have used conservative scores and would expect a very large propotion of our assignments to be correct. However, given that we deal with over two hundred sequences, which together have about a thousand domains, we might also expect that a few assignments will be incorrect, and that some assignments will be missed because of the limitations of some of the hidden Markov models.

The criticisms made by Hobert et al. (Hobert et al., 2004) do not affect the novel and significant parts of our paper. We show that about half of the IgSF proteins in *C. elegans* and three-quarters of those in *Drosophila* have evolved since the divergence of the two organisms. The larger size of the *Drosophila* IgSF repertoire involves mainly cell surface and secreted proteins, and many of these have arisen through gene duplications. We believe that this overall expansion of the IgSF must be one of the factors that contributed to the formation of the more complex physiology of *Drosophila*. It is difficult to understand the assertion made by Hobert et al. (Hobert et al., 2004) that this view is invalidated by the increases in the repertoire produced by the alternative splicing of genes. Both factors are clearly important. In fact, the protein they take to illustrate the importance of splicing, DSCAM, is also a good example of repertoire expansion: there are probably four DSCAM sequences in *Drosophila* and none in *C. elegans.*

References

Aurelio, O., Hall, D. H. and Hobert, O. (2002). Immunoglobulin-domain proteins required for maintenance of ventral nerve cord organization. *Science* **295**, 686-690.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755-763.

Flaherty, D. B., Gernert, K. M., Shmeleva, N., Tang, X., Mercer, K. B., Borodovsky, M. and Benian, G. M. (2002). Titins in C. elegans with unusual features: coiled-coil domains, novel regulation of kinase activity and two new possible elastic regions. *J. Mol. Biol.* **323**, 533-549.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903-919.

Hegyi, H. and Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632-1640.

Hobert, O., Hutter, H. and Hynes, R. (2004). The immunoglobulin superfamily in *Caenorhabditis elegans* and *Drosophila melanogaster. Development* **131**, 2237-2238.

Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C. B., Mastwal, S., Zhu, X. P., Scheel, J. et al. (2000). Cell biology: conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989-994.

Hynes, R. O. and Zhao, Q. (2000). The evolution of cell adhesion. *J. Cell Biol.* **150**, F89-F96.

Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846-856.

Kubiseski, T. J., Culotti, J. and Pawson, T. (2003). Functional analysis of the Caenorhabditis elegans UNC-73B PH domain demonstrates a role in

activation of the Rac GTPase in vitro and axon guidance in vivo. *Mol. Cell. Biol.* **23**, 6823-6835.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P. and Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucl. Acids Res.* **32**, D142-D144.

Madera, M. and Gough, J. (2002). A comparison of profile hidden Markov model procedures for remote homology detection. *Nucl. Acids Res.* **19**, 30.

Popovici, C., Isnardon, D., Birnbaum, D. and Roubin, R. (2002). Caenorhabditis elegans receptors related to mammalian vascular endothelial growth factor receptors are expressed in neural cells. *Neurosci. Lett.* **329**, 116-120.

Roy, P. J., Zheng, H., Warren, C. E. and Culotti, J. G. (2000). mab-20 encodes Semaphorin-2a and is required to prevent ectopic cell contacts during epidermal morphogenesis in Caenorhabditis elegans. *Development* **127**, 755-767.

Teichmann, S. A. and Chothia, C. (2000). Immunoglobulin superfamily proteins in *Caenorhabditis elegans*. *J. Mol. Biol.* **296**, 1367-1383.

Vogel, C., Teichmann, S. A. and Chothia, C. (2003). The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. *Development* **130**, 6317-6328.

Christine Vogel*, Sarah A. Teichmann and Cyrus Chothia

MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK

*Author for correspondence (e-mail: cvogel@mrc-lmb.cam.ac.uk)