

The Evolution of Cell Adhesion[Ⓢ]

Richard O. Hynes* and Qi Zhao[‡]

*Howard Hughes Medical Institute, Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 20139; and [‡]Celera Genomics, Rockville, Maryland 20850

The Essence of Being Metazoan

Multicellular organisms clearly require mechanisms for intercellular communication and, perhaps even more basically, for intercellular cohesion. The most primitive sponges and coelenterates depend on cell adhesion for their organismal organization; so do insects, nematodes and vertebrates. What molecules and mechanisms are common among these different phyla and which ones differ and why?

The availability of the euchromatic genomic sequences of *Drosophila melanogaster* (Adams et al., 2000; Rubin et al., 2000; <http://www.celera.com>) and *Caenorhabditis elegans* (http://www.sanger.ac.uk/Projects/C_elegans/) makes it possible to address these questions with much more confidence than heretofore. We searched the *Drosophila* (and, to a lesser extent, the *C. elegans*) genomic sequences using a large number of vertebrate sequences of adhesion proteins. We also conducted searches for particular domains prevalent in adhesion proteins (Kreis and Vale, 1999) and made extensive use of the listings of *Drosophila* transcripts sorted by domain family that are available at the EBI web site (<http://www.ebi.ac.uk/proteome/>). Because of the complex, multi-domain nature of many adhesion proteins (see <http://expsy.cbr.nrc.ca/cgi-bin/lists?extradom.txt> and Bork and Bairoch [1995] or Kreis and Vale [1999] for listings), significant matches were frequently obtained because of the presence of some shared domains, while other domains were missing. So, homologues were further analyzed for their domain complement and arrangement (using Pfam and Interpro) and for extent of homology by pairwise Blast comparisons. For *C. elegans* homologues, we referred frequently to the detailed analysis presented by Hutter et al. (2000; <http://www.mpimf-heidelberg.mpg.de/ewgdn/>). When true orthologues appeared to be absent from flies or worms, we searched extensively with individual domains and with unique segments against the entire genomic sequences. Naturally, all statements about absence of particular genes must be qualified by several cautions. First, the sequence has some gaps and some genes do exist in the heterochromatin, most of which remains unanalyzed. Second, it is always possible that some genes are missed during curation or that distant homologues

might be missed in the search and further refinement of these analyses may reveal additional genes. However, we are fairly confident in our claims that certain genes and domains are absent.

We review here the results of our analyses and discuss some of the implications. Overall, we identified ~500 *Drosophila* genes that are candidates for involvement in cell adhesion (~4% of the genome). The molecules mediating cell–cell and cell–matrix adhesion exemplify both extreme conservation among diverse organisms and considerable diversification in different phyla, presumably to meet different biological needs.

Cell–Cell Adhesion

Many of the major classes of cell–cell adhesion molecules were already known to be shared among vertebrates and invertebrates and the genome sequences confirm this picture in great detail. However, they also reveal some interesting differences.

The two major classical groups of cell–cell adhesion receptors, cadherins and immunoglobulin superfamily (Ig-SF)¹ proteins, are both well represented in *Drosophila*. We found 17 convincing cadherin homologues, five of which were previously known (Fig. 1; Table S1, all supplemental tables S1–S8 are available at <http://www.jcb.org/cgi/content/full/150/2/F89/DC1>). As the cadherin superfamily has grown, the nomenclature has become somewhat confused. In this article we will define “classical cadherins” by their cytoplasmic domain homology and restrict use of protocadherin to refer to homologues of the clustered “protocadherins” (or CNRs) of vertebrates (Tepass, 1999; Yagi and Takeichi, 2000). Searches for cytoplasmic domain matches in *Drosophila* revealed three classical cadherins containing catenin-binding segments (DE-cadherin, DN-cadherin, and a novel cadherin closely related, and closely linked, to DN-cadherin). In contrast, no matches were found with the conserved vertebrate protocadherin cytoplasmic domains (see below). There are

[Ⓢ]The online version of this article contains supplemental material.

Address correspondence to Richard Hynes, E17-227, Center for Cancer Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 20139. Tel.: (617) 253-6422. Fax: (617) 253-8357. E-mail: rohynes@mit.edu

¹Abbreviations used in this paper: ADAM, disintegrin metalloproteinase; CNR, cadherin-related neural receptor; CT, COOH-terminal; ECM, extracellular matrix; FB-C, fibrinogen COOH-terminal domain; Fn3, fibronectin type III repeat; Ig-SF, immunoglobulin superfamily; LDLR, LDL receptor; LM, laminin; LM-G, laminin G domain; LRR, leucine-rich repeat; PI, phosphatidylinositol-linked; PS, position-specific; RGD, arginine-glycine-aspartate; TB, TGFβ-binding domain; TM, transmembrane; TM4, tetraspanin; TSP-1, thrombospondin type 1 domain; vWA, von Willebrand A domain; vWD, von Willebrand D domain.

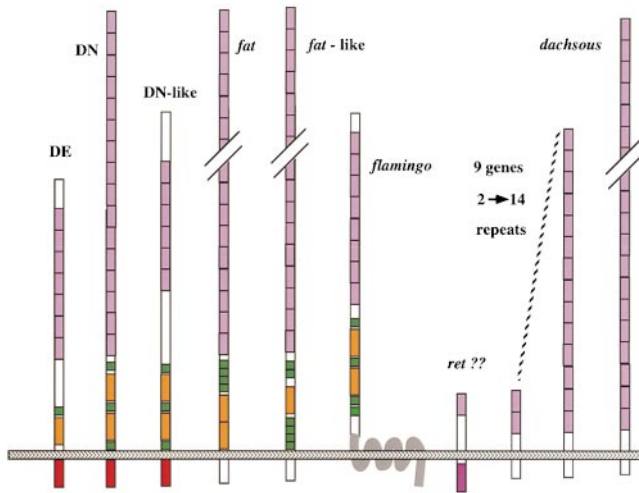


Figure 1. The cadherin superfamily in *D. melanogaster*. Three cadherins have homology with the catenin-binding cytoplasmic domains (red) of vertebrate cadherins; two previously known (DE and DN) and one novel (DN-like). All three contain EGF (green) and LM-G (orange) domains in addition to cadherin repeats (lilac). *Fat*, *dachsous* and *flamingo* were previously known and there is a novel *fat*-like homologue. 10 additional cadherin genes encode proteins of varying sizes (1–14 cadherin repeats and no EGF or LM-G repeats) but with no obvious cytoplasmic domain homology with vertebrate cadherin superfamily members. One of these genes has a protein kinase domain and bears some similarity with the *ret* oncogene, which also has weak cadherin homology. For CT numbers, and other information, see Tables S1 and S6.

several known large cadherin homologues; *fat*, *dachsous*, and *flamingo/starry night*, which has a secretin receptor-type seven transmembrane segment, as well as a novel homologue of *fat* (Fig. 1). Homologues of these large cadherins exist in both vertebrates and nematodes, as do classical cadherins. Most of these cadherin homologues contain EGF and LM-G repeats along with cadherin repeats and none of them really fits the mold of classical vertebrate cadherins (i.e., five extracellular cadherin repeats and a catenin-binding cytoplasmic domain). In addition there are 10 other cadherins with varying numbers (1–14) of cadherin repeats and no obvious matches with the conserved cytoplasmic domains of vertebrate cadherins or protocadherins (see Fig. 1; Table S1).

Thus, *Drosophila* and *Caenorhabditis* have similar numbers and spectrum of cadherin homologues (17 and 13, respectively), but vertebrates have many more. Clearly this family of Ca^{++} -dependent cell–cell adhesion molecules arose early in metazoan evolution and evolved early into several distinct variant subtypes (classical, *fat*-like, and *flamingo*-like) that are conserved to this day. Additional subtypes (protocadherins, desmocollins, desmogleins) arose later in chordates but not in the two sequenced invertebrates (see also below).

The Ig-SF of adhesion receptors is larger than the cadherin superfamily in all three phyla. *Drosophila* has ~150 genes containing Ig domains (more than *Caenorhab-*

ditis, which has ~70). They can be sorted roughly into several groups (Tables S2–S4). There are around 50 Ig-SF genes encoding 1–2 Ig domains, most without obvious transmembrane (TM) domains (Table S2). They could be involved in cell adhesion or, as secreted proteins, may participate in intercellular communication or in binding to pathogens. A second group has three or more (up to nine) Ig domains but no other recognizable domains. Many of these, but not all, have predicted TM domains and are likely involved in cell adhesion (Table S3). A third group contains one or more Ig domains in tandem with other domains; EGF, TSP-1, LRR, collagen, sushi, and, most frequently, Fn3 domains (Table S4, A and B). These are likely to be (or known to be) involved in cell adhesion or as receptors for ligands such as netrins (e.g., CT20824, unc5-like) and homologues with similar structures are known in nematodes and vertebrates. Finally, several Ig-alone proteins have associated protein tyrosine kinase domains and are all presumably signaling proteins (Table S4 A).

In addition to their presence in Ig/Fn3 adhesion receptors (Table S4 B), Fn3 domains also exist in around a dozen other genes (Table S5). Some are clearly signaling receptors with tyrosine kinase or tyrosine phosphatase domains. Similar Ig and/or Fn3 kinase and phosphatase receptors exist in *C. elegans* and in vertebrates. The other *Drosophila* Fn3 proteins (Table S5) are presumably adhesion receptors or ECM molecules. Interestingly, Fn3 repeats do not appear in tandem arrays with EGF or other disulfide-bonded domains as is common in vertebrate ECM molecules (see more below). Also absent from *Drosophila* are the extremely repeated Fn3 (myotactin) or Ig (hemicentin) domain proteins found in *C. elegans*. (Note: Ig and Fn3 domains also exist in all three phyla in large intracellular muscle proteins such as titin, twitchin and projectin, presumably not involved in cell adhesion and beyond the scope of this article).

Thus, *Drosophila*, in common with nematodes and vertebrates, makes extensive use of Ig and/or Fn3 domains for cell surface adhesion and/or signaling receptors. Like cadherins, these cell interaction molecules must have arisen early in metazoan evolution and diverged to perform differing functions before separation of the arthropod, nematode, and deuterostome lineages, and have been conserved since.

Other proteins likely to be involved in cell adhesion and shared among the three phyla include many EGF family proteins; >100 in each of *D. melanogaster* (Table S6) and *C. elegans*, leucine-rich repeat (LRR) proteins (~50 in *C. elegans* and twice as many in *D. melanogaster*; see Table S7) and C-type lectins of which the worm has many more (165) than the fly (37) (see <http://www.ebi.ac.uk/proteome/> and <http://www.mpimf-heidelberg.mpg.de/ewgdn/> for lists). There are 32 TM4-superfamily (tetraspanin) proteins in *D. melanogaster*, 11 of them linked in a cluster on chromosome 2R (see <http://www.ebi.ac.uk/proteome/>), and 7 ADAMs (disintegrin-metalloproteinase) family members (see Tables S6 and S7). Again, all of these families must have evolved early, before divergence of the three lineages. It appears that once cells evolved some basic mechanisms for sticking together sensibly they did not let go, either of each other or of those adhesion receptors.

Cell–Matrix Adhesion

The second arm of cell adhesion, attachment to basement membranes, appears equally ancient and also exquisitely conserved. Coelenterates have basement membranes, as do all more complex animals. The basic constituents of basement membranes are type IV collagen, laminin, nidogen/entactin and proteoglycans of the perlecan type; these molecules are all highly conserved (Fig. 2). *Drosophila* has a laminin comprising three subunits (α , β , γ ; formerly A, B1, B2). These have been known for some time and are clearly related to laminins in vertebrates (of which there are many) and in *C. elegans* (which has 2α 's, 1β and 1γ). *D. melanogaster*, like *C. elegans*, has a second alpha subunit. Similarly, both the fly and the worm have a single pair of type IV collagen genes. In vertebrates, which have three pairs, it is notable that each pair is organized in an antiparallel head-to-head arrangement with a common promoter region between (Kreis and Vale, 1999). It is striking to find that the two *Drosophila* type IV collagen genes are also closely linked head-to-head on chromosome 2L (see Fig. 2). It will be of interest to determine whether they are also regulated by a common promoter.

Two other common constituents of vertebrate basement membranes are the proteoglycan perlecan and the glyco-

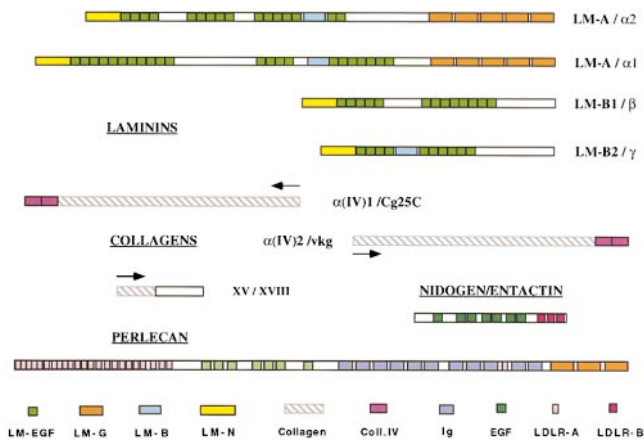


Figure 2. Basement membrane proteins of *D. melanogaster*. The three laminin subunits (A/ α 1, B1/ β , and B2/ γ) were described some time ago; A/ α 2 more recently (see Table S6 for details). All are homologous with vertebrate laminins and it is notable that *C. elegans* has a very similar complement of four laminin subunits (Hutter et al., 2000). The set of three conserved collagen genes (two α IV subunits and one α XV/XVIII homologue) is also replicated in *C. elegans*. In *D. melanogaster*, the two type IV collagen genes (CT12803 and CT25584) are linked in head-to-head organization as are all three type IV collagen pairs in vertebrates. The two type IV collagen genes in *C. elegans* are not linked. Several other genes in *D. melanogaster* contain short collagen segments but there are no fibrillar collagens with long collagen triple helices. The *D. melanogaster* homologues of nidogen/entactin (CT32053) and perlecan (CT23996) have similar domain compositions (see this figure and Table S6) as their homologues in *C. elegans* and vertebrates, although the numbers of repeats vary among phyla. The set of genes shown in this figure has been well conserved since some early common metazoan ancestor of arthropods, nematodes, and vertebrates developed a basement membrane.

protein nidogen/entactin. Both have homologues in *C. elegans* and *D. melanogaster*. The perlecan of worms, flies and vertebrates all comprise tandem arrays of LDLR-A, LM-EGF, Ig and LM-G domains (see Fig. 2; Hutter et al., 2000) but the numbers of repeats vary somewhat. Similarly the structures of the nidogen/entactin homologues vary, although each has EGF and LDLR-B domains (see Fig. 2 and Hutter et al., 2000; Kreis and Vale, 1999).

Therefore, it seems clear that these four complex proteins; type IV collagen, laminin, nidogen/entactin, and perlecan, formed the basis of an early basement membrane that has been preserved in molecular detail ever since.

Several other ECM proteins are well conserved among the three phyla, including collagen XV/XVIII (CT14872, see Fig. 2), SPARC/osteonectin (CT19876), netrins (CT27014 and CT29512, see Table S6), and the anosmin/Kallmann syndrome protein (CT19368, see Table S4). All these were first identified in vertebrates and have good homologues in *C. elegans* as well as *Drosophila*. Netrins are well established neural guidance molecules as are slit (CT21700 and CT37068, see Table S6) and semaphorins (see Table S7), which also occur in all three phyla. The functions of collagen XV/XVIII, SPARC/osteonectin, and anosmin/Kallmann syndrome protein are less clear but, given their strong evolutionary conservation, are likely to be fundamental and well worthy of further study. As we will discuss below, many other ECM molecules show much less conservation.

Extracellular matrix is clearly important but equally significant are the receptors by which cells attach to ECM; these too are well conserved. The major ECM receptors are integrins, $\alpha\beta$ heterodimeric receptors linking ECM to the cytoskeleton (Hynes, 1992). They are found in organisms ranging from sponges, corals, nematodes and echinoderms to mammals (Burke, 1999). *Drosophila* has two β subunits (β PS and β v, CT40473 and CT5192, both previously known) and five α subunits of which three (α PS1-3) were known (Gotwals et al., 1994b; Stark et al., 1997; Grotewiel et al., 1998; Fig. 3). The two novel α subunits are most closely related to α PS3 and one (which we call α PS4) is closely linked to α PS3 (chromosome 2R, 5IE-F). α PS5 is also on 2R although not so closely linked (59E) but it is also similar in structure. Given this homology, it is likely that all five α subunits complex with β PS to form five PS integrins (already known for α PS1-3); β v so far has no known α partner. It is clear that α PS1 β PS and α PS2 β PS are, respectively, receptors for laminin and RGD-containing ECM proteins (Zavortink et al., 1993; Gotwals et al., 1994a,b) and each of these two α subunits is most homologous with a set of functionally related vertebrate α subunits (laminin-specific; α 3, α 6, α 7, α PS1 or RGD-specific; α 5, α 8, α v, α IIB, α PS2; see Fig. 2). It is notable that *C. elegans* also has an orthologue of each of these subfamilies (Gotwals et al., 1994b), F54G8.3 (*ina-13*) and F54F2.1 (*pat-2*), respectively.

It seems evident that some early metazoan evolved two integrins, one laminin-specific and one recognizing RGD or something like it, and these two families have been preserved ever since. Since α PS1 β PS and α PS2 β PS are frequently expressed by apposed tissues separated by extracellular matrix (Fristrom et al., 1993; Brower et al., 1995), it is an intriguing hypothesis that the two classes of inte-

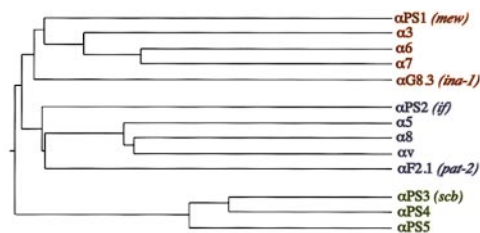


Figure 3. The integrin α subunits of *D. melanogaster*. There are five integrin α subunits in *Drosophila*. The figure shows a dendrogram developed using AlignX and demonstrates that α PS1 (CT5254) is most closely related to a set of vertebrate integrins (α 3, α 6, α 7) that bind laminin as does α PS1. One of the two *C. elegans* α subunits (α G8.3) appears to be an orthologue of this set. α PS2 (CT27194) is most closely related to RGD-binding vertebrate integrins (α 5, α 8, α v, α IIb) and is known to bind the RGD proteins; again there is a worm orthologue of this group (α F2.1). These two distinct classes of integrin clearly arose very early in metazoan evolution. In contrast, α PS3 (CT21280/CT36929; Stark et al., 1997; Grotewiel et al., 1998) and two novel subunits, α PS4 (CT37393) and α PS5 (CT17068), form a separate evolutionary group with no obvious orthologues in other species. These three genes are all on 2R and α PS3 and α PS4 are very closely linked. They may be a more recent evolutionary development. Gene names are given in italics.

grins might originally have evolved to attach two different cell layers to opposite sides of a basement membrane (e.g., in a simple two-layered organism such as hydra). *C. elegans* still makes do with just two integrins, while *Drosophila* has several additional integrins (of so far unknown specificity). It is of some interest that *Drosophila* has evolved a small family of its own integrins (α PS3, α PS4, and α PS5) not closely homologous with orthologues in other phyla. A similar phenomenon has been noted before for echinoderm integrin β subunits (Burke, 1999). Are these integrins specialized for specific fly or sea urchin adhesive functions? If so, what are they? It is of some interest that mutations in α PS3 affect short-term memory in flies (Grotewiel et al., 1998). It is unknown whether or not vertebrate integrins may play a similar role. Vertebrates meanwhile have evolved many more integrins (8 β subunits and 18 α subunits known to date). Around half of the vertebrate α subunits include an extra inserted I domain (homologous with von Willebrand A domains). I domains are found in many integrins that bind to collagens and in leukocyte integrins but no I domains occur in fly or worm integrin α subunits. Indeed, we could not detect vWF-A domain homologues in *Drosophila* adhesion molecules except, notably, for β v. Perhaps β v functions alone or as a homodimer. We will return later to the issue of differential evolution of adhesion molecules.

Cytoskeletal Connections

A key feature of cell adhesion is the linkage of cell adhesion receptors to the cytoskeleton. This affects not only the intracellular consequences of cell adhesion (cell shape and polarity, cytoplasmic organization and cell motility) but also intracellular signal transduction and even the efficacy of the adhesive interactions at the extracellular sur-

face. The cytoskeletal connections of cadherins and integrins have been extensively studied in vertebrates and appear to be conserved in many details in *Drosophila*, although some key features appear to be absent.

Classical vertebrate cadherins link via β -catenin to α -catenin and thence to the actin cytoskeleton. The fly homologue of β -catenin is *armadillo* (CG 11579; three alternatively spliced forms). Two β -catenin-like molecules are known in vertebrates (β -catenin and plakoglobin or γ -catenin). *Drosophila* also has a homologue of vertebrate α -catenins (CT39986). Thus, this cytoskeletal connection is well conserved.

In contrast with this conservation of classical catenin-binding via cadherins, other classes of cadherin known in vertebrates are missing. Desmocollins and desmogleins are cadherin homologues found in vertebrate desmosomes. They have characteristic cytoplasmic domains that link via desmoplakins to intermediate filaments. Since *Drosophila* lacks intermediate filaments (Goldstein and Gunawardena, 2000) it is perhaps not surprising that we could find no convincing homologies for the characteristic cytoplasmic domains of desmocollins and desmogleins. More surprising is that we also could not find them in *C. elegans*, which does have intermediate filaments. *Drosophila* and *C. elegans* also lack the β 4 integrin subunit that is linked to intermediate filaments in vertebrates.

The more typical integrin-actin microfilament connection is well conserved in *Drosophila*, which has single copy genes for the cytoskeleton linker/adaptor proteins of integrins; talin, α -actinin, vinculin, paxillin, tensin, as well as the integrin-linked signal transduction molecules, FAK, ILK, p95PKL and p130CAS (Fig. 4). Many of these proteins occur in multiple copies in vertebrates. Their occurrence as single genes in *Drosophila* (and *C. elegans*?) will facilitate genetic and other analyses of their functions in this evidently ancient ECM-integrin-cytoskeleton connection.

Another well analyzed transmembrane ECM-cytoskeleton linkage is the laminin-dystroglycan/sarcoglycan-dystrophin linkage. There are single *Drosophila* homologues of dystroglycan (CT41273) and γ/δ sarcoglycan (CT34621), two transmembrane proteins linking laminin to dystrophin in vertebrates. As mentioned above, laminin exists in *Drosophila*, as does dystrophin together with dystrobrevin and syntrophins (Goldstein and Gunawardena, 2000). The dystroglycan/sarcoglycan complex appears to be simpler in *Drosophila*, which again may make it easier to analyze.

Variations on Basic Themes

In contrast with the high degree of evolutionary conservation in cell-cell and cell-matrix adhesion discussed above, other aspects of cell adhesion and, in particular, extracellular matrix proteins show considerable variation among flies, worms and vertebrates. We have already mentioned the abundance of C-type lectins in nematodes as compared with fruit flies (other lectins are also very numerous in *C. elegans*). Why is this? *Drosophila* has made much more use of Ig and LRR domains than has *C. elegans*. Again, why?

Both species have elaborated large, complex, extracellular matrix molecules. It is far from clear what the advan-

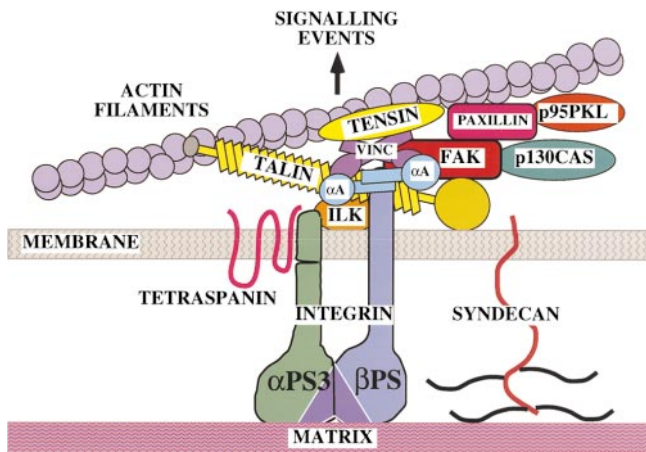


Figure 4. Focal contact constituents in *D. melanogaster*. The diagram depicts an integrin-mediated transmembrane linkage between extracellular matrix and the actin cytoskeleton. Single genes encode talin (CT21159), α -actinin (CT14163 and CT14232), vinculin (CT11081), tensin (CT26639), paxillin (CG18576; Wheeler, G.N., and R.O. Hynes, manuscript submitted for publication), p130CAS (CT1293), FAK (CT28129), ILK (CT29478), and syndecan (CT27296). There are multiple genes for integrin subunits and tetraspanins. The arrangement shown is based on data from vertebrate homologues, not in most cases, on data from *Drosophila*. However, the good conservation of all these associates of integrins strongly suggests that the cytoskeletal organization and signaling functions of these adhesion receptors are conserved.

tage of very large ECM molecules might be. One can clearly make stable polymers from small proteins (intermediate filaments, bacterial flagellae) so structural arguments are not in general terms compelling. Even for some of the best understood ECM proteins, we can only assign functions to a small fraction of the repeated domains. Yet the others are equally well conserved. What are they doing? Clearly, many of the domains that have been used to elaborate matrix and other adhesive proteins are good at binding other proteins; that is what many of the well defined domains do. Presumably the others do something similar. They may bind other ECM proteins. They may engage multiple cell surface receptors to trigger complex intracellular responses, and, as discussed earlier, there are certainly a large number of likely adhesion receptors with unassigned functions. Another possibility is that ECM proteins act as docking sites for diffusible factors. That is known to happen and may be more prevalent than we know. Classical mathematical models that attempt to explain morphogenetic gradients typically invoke both freely diffusible and more stably anchored gradients of morphogens. Binding to ECM proteins could well be one way to establish the more stable, slowly changing gradients. Perhaps that is what explains the exuberant elaboration of domains in many ECM proteins. In the absence of a clear explanation for the multiplicity of domains in any one matrix protein, it is even harder to understand why *C. elegans* should proliferate Fn3 domains in myotactin or Ig domains in hemiceptin, while *D. melanogaster* concatenates von Willebrand D, trypsin inhibitor and other domains in hemolactin (CT21553) and why both species have proteins with multiply repeated EGF and CUB domains (Table S6).

It is a little easier to offer rationalizations for the extreme expansion of the set of collagen genes in *C. elegans*. In contrast with the rather limited set of collagens found in *Drosophila*, *C. elegans* has around 170 collagen genes, many of them encoding cuticular collagens. The collagenous cuticle provides an exoskeleton for *C. elegans*. Vertebrates have also used a wide variety of collagens, particularly extended fibrillar collagens to construct endoskeletons (cartilage, bones) and the connections to them (tendons) as well as the interstitial connective tissue that provides structural strength to vertebrate tissues. Neither flies nor nematodes appear to have elaborated such fibrillar collagens. Indeed, apart from the basement membrane collagens mentioned earlier, *Drosophila* has only a few genes encoding short collagen segments.

Has *Drosophila* evolved ECM proteins specialized, for example, to attach muscle cells to the chitin exoskeleton? A number of *Drosophila* ECM proteins are concentrated at muscle attachment sites. One of these, tiggerin is composed of 16 repeats of 75 ± 2 amino acids (CT36389; Fogerty et al., 1994). It has no clear homologues in vertebrates or nematodes. This may be one example of a phenomenon common in the *C. elegans* genome; namely, the elaboration of repeated domains that are, so far, largely specific to nematodes (Hutter et al., 2000). Whether the same is widely true in *Drosophila* is not yet clear but could be revealed by appropriate analyses of the proteomic sequence. Several other *Drosophila* ECM proteins that have been described as being concentrated at muscle attachment sites or at sites of apposition of the two surfaces of wings, are generic, such as laminin, or at least have homologues in vertebrates. Examples include ten-a and ten-m, two EGF repeat proteins, as well as peroxidasin, vanin, and glutactin (see Table S8).

What's Missing and Why?

We have already mentioned the striking absence of fibrillar collagens and of intermediate filaments in *Drosophila*. It must be noted that ablation of intermediate filaments from vertebrate cells frequently has remarkably subtle cellular effects; the defects lie rather at the tissue structural level (e.g., skin blisters). Perhaps flies do not need the mechanical strength provided by IF and fibrillar collagens because of the existence of a chitinous exoskeleton.

Many well known vertebrate ECM proteins appear to be missing from the *Drosophila* genome. These include fibronectin, vitronectin, elastin, fibulins, osteopontin, von Willebrand factor, thrombospondins, tenascins, and fibrinogen. In many cases some of the characteristic domains are present, such as the Fn3 domains characteristic of fibronectin and tenascin, and vWF-C and D domains. However, in each of these examples, other domains are missing. We were unable to detect any FN type I and II domains in the *Drosophila* or *C. elegans* genomic sequences and only a few vWF-A domains, none associated with other vWF domains. Although both EGF and Fn3 domains are prevalent in *Drosophila*, we could not detect any genes that contained these two domains together (as in tenascins) although that had been claimed for ten-m (Baumgartner et al., 1994). Vertebrate thrombospondins are assembled from TSP-1, TSP-2 (EGF), and TSP-3 do-

mains. Although TSP-1 and EGF domains do occur in *Drosophila* (Tables S6 and S7), we did not find them together in one gene and found no TSP-3 domains. Fibrinogen COOH-terminal domains are present in 10 genes in *Drosophila* but typically in relatively small proteins, where it is known that they function as pathogen-binding domains. They do not appear to assemble into large coiled-coil-containing molecules like vertebrate fibrinogen.

Many of these vertebrate ECM proteins are not essential for life (Hynes, 1996). Mice lacking vitronectin, osteopontin, fibrinogen, or von Willebrand factor are all viable, although the last two have bleeding problems. One interpretation is that they subservise specialized functions (e.g., hemostasis, wound repair). However, that cannot be said of fibronectin; ablation of the fibronectin gene produces early embryonic lethality. What could be so important about a gene that is absent in fruit flies and nematodes? One plausible hypothesis is that fibronectin is essential for blood vessel formation. Indeed, the lethal FN-null phenotype includes a major early vascular defect (George et al., 1993). Other vertebrate genes, mutations in which yield vascular defects, include fibrillar collagens, elastin and fibrillins. We could not identify convincing homologues of any of them. The few weak matches with elastin appeared to be only alanine-rich stretches and, although EGF repeats are common, we did not find them in tandem with the TGF β -binding domains characteristic of fibrillins.

It seems that an entire set of genes necessary to construct blood vessels and contain the pressure of circulating blood was elaborated during the evolution of vertebrates. This entailed novel assemblages of ancient domains (Fn3, EGF, vWA, vWD, TSP-1) and the development of new domains (e.g., Fn1, Fn2, TB) and novel proteins (e.g., elastin). Several of the other missing proteins (vitronectin, fibrinogen, von Willebrand factor) are prevalent in, or unique to, blood. Further pursuit of this idea led us to search for VEGFs and angiopoietins. We found no convincing matches for VEGF and, although the fibrinogen C domain proteins somewhat resemble angiopoietins, the homology is limited to the FB-C domain. Furthermore, we have found no tyrosine kinase receptors with Ig, EGF and Fn3 domains like the tie2 receptor for angiopoietins. It appears that these crucial genes involved in vascular development in vertebrates are absent from *Drosophila*. There are parallels between vascular and insect tracheal development (Samakovlis et al., 1996) but vascular development clearly involves additional sets of genes.

Returning to the earlier discussion of additional vertebrate integrins and the paucity of integrin I and vWF-A domains in *Drosophila* and *C. elegans*, it is worth noting that several A domains in vWF bind collagen and that I

domains bind collagen in several vertebrate integrins. Perhaps the elaboration of A/I domains in vertebrates accompanied the proliferation of collagen genes. Other I domain integrins are selectively expressed on white blood cells. Those integrins are absent from the two invertebrates, as are selectins, another class of adhesion receptors involved in adhesion of white blood cells. Selectins rely on a C-type lectin plus an EGF domain and, while C-type lectins are present in both *C. elegans* and *Drosophila*, we found no CL/EGF pairs in the fruit fly. Genes labeled selectin-like in the worm and fly lack EGF domains, although some fly genes do contain C-type lectins together with sushi domains, also found in selectins. Thus, two major classes of receptors key in adhesion functions of blood cells appear to be vertebrate inventions.

It appears to us that a large number of genes involved in the development, maintenance and function of the vasculature in vertebrates evolved only in the chordate lineage (Table I). This is not a great surprise but it points to the value of further genomic and genetic analyses of vertebrate systems.

Another organ system that is much more elaborate in vertebrates is the nervous system and it appears that, there too, vertebrates have elaborated genes encoding adhesion molecules that are not found in *Drosophila* or *C. elegans*. Vertebrates have many more cadherins than do flies and worms, and many of them are expressed in the brain (Yagi and Takeichi, 2000). The large family of protocadherins encoded by complex genetic loci, each comprising a set of homologous extracellular and transmembrane segments that become linked to a common cytoplasmic domain (by alternative splicing or, perhaps, DNA rearrangement), are also prevalent in the brain. Flies and worms lack the cytoplasmic domain sequences characteristic of these protocadherins. The vertebrate protocadherins have been implicated in synaptic function and as receptors for reelin, an ECM protein which affects neuronal migration during cortical and cerebellar migration. We were unable to detect any homology with reelin in the *Drosophila* genome. We also could not detect a good homologue of agrin, a protein involved in neuromuscular junction organization. Although genes do occur that contain most of the domains found in agrin, no one gene contains them all. So it appears that, although vertebrates and invertebrates share many common adhesive molecules that guide neuronal development (e.g., netrins, semaphorins, slit, and their receptors, Eph receptor and ephrins), vertebrates have developed additional adhesion receptors (e.g., protocadherins) and ligands (e.g., reelin, agrin) that perform important functions in development of the nervous system that may be specific to vertebrates.

In conclusion, although the analysis that has been possi-

Table I. Vascular Proteins Absent from *Drosophila melanogaster*

Vascular development	Vascular integrity	Blood proteins	Vascular adhesion receptors
Fibronectin	von Willebrand factor	Fibronectin	Selectins
Thrombospondins	Fibrinogen	Vitronectin	β 2 Integrins
Elastin	Fibrillins	Fibrinogen	
VEGF	Elastin	von Willebrand factor	
Angiopoietins/tie 2	Fibrillar collagens	Thrombospondins	

ble to date represents only the beginning in extracting insights from comparative analyses of the genomes of flies, worms and vertebrates, some clear messages are apparent. They are not particularly surprising in outline but the details are stimulating and informative and the fact that one is able to look at essentially the entire blueprint for the organism adds strength to the hypotheses that can be formulated and the sequences open the route to testing those hypotheses.

Examination of the set of genes encoding adhesion proteins reveals both the great conservation of some basic processes as well as the elaboration of new genes and processes during evolution. The detailed molecular conservation of basic cell-cell and cell-matrix adhesions and of basement membrane structure is remarkable and confirms yet again the evolution from a common ancestor of arthropods, nematodes and mammals.

For metazoans to evolve from single cells they had to invent cell adhesion. This apparently involved evolution of new protein domains. Ig, EGF, TSP-1, LDLR-A, C-type lectins, cadherins, and collagen triple-helix domains are all absent from yeast, as are laminins, tyrosine kinases, integrins, band 4.1 proteins and many others involved in cell-cell interactions. But, once these domains and genes evolved, they have been used over and over again. The complex proteins elaborated early in metazoan evolution to assemble basement membranes and attach cells to them and to one another have been conserved in great detail ever since. Many adhesive routines appear to be the same in flies, worms and people, although often duplicated and replicated in vertebrates. Such processes can be very effectively analyzed in invertebrates. However, it is equally clear from browsing the genomes and proteomes that vertebrates have evolved some new tricks not found in flies and worms. In many cases new proteins have been assembled from new arrangements of old domains. However, new domains and entirely new proteins have also evolved that have no close counterparts in invertebrates. This is particularly evident in vascular biology and in some aspects of neurobiology and is likely to be true for some other uniquely vertebrate functions such as neural crest migration. It will be fascinating to be able to look at the entire set of human genes in the near future and ask what additional new adhesive tricks have been elaborated during our evolution from the common ancestor of protostomes and deuterostomes.

The genomic analysis of *Drosophila* reinforces the conclusion that it and *C. elegans* are wonderful models for some aspects of vertebrate life but it also shows that mice and zebrafish and the human genomic sequence will offer insights that we cannot hope to gain from invertebrates. That is particularly so for the multicellular processes in which cell adhesion plays an important part.

We thank Rolf Apweiler, Larry Goldstein, Tom Maniatis, and Masatoshi Takeichi for helpful inputs during the analysis, Denisa Wagner for critical

review of the manuscript and Charlie Whittaker for help with Fig. 3. Richard Hynes is an Investigator of the Howard Hughes Medical Institute.

Submitted: 1 June 2000

Revised: 27 June 2000

Accepted: 27 June 2000

References

- Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides et al., and J.C. Venter. 2000. The genome sequence of *Drosophila melanogaster*. *Science*. 287:2185–2195.
- Baumgartner, S., D. Martin, C. Hagios, and R. Chiquet-Ehrismann. 1994. *ten^m*, a *Drosophila* gene related to tenascin, is a new pair-rule gene. *EMBO (Eur. Mol. Biol. Organ.) J.* 13:3728–3740.
- Bork, P., and A. Bairoch. 1995. Extracellular protein modules: a proposed nomenclature. *Trends Biochem. Sci.* 20(Special Poster Supplement): TIBSC02.
- Brower, D.L., T.A. Bunch, L. Mukai, T.E. Adamson, M. Wehrli, S. Lam, E. Friedlander, C.E. Roote, and S. Zusman. 1995. Nonequivalent requirements for PS1 and PS2 integrins at cell attachments in *Drosophila*: genetic analysis of the α PS1 integrin subunit. *Development*. 121:1311–1320.
- Burke, R.D. 1999. Invertebrate integrins: structure, function and evolution. *Int. Rev. Cytol.* 191:257–284.
- Fogerty, F.J., L.I. Fessler, T.A. Bunch, Y. Yaron, C.G. Parker, R.E. Nelson, D.L. Brower, D. Gullberg, and J.H. Fessler. 1994. Tiggriin, a novel *Drosophila* extracellular matrix protein that functions as a ligand for *Drosophila* α PS2 β PS integrins. *Development*. 120:1747–1758.
- Fristrom, D., M.E. Wilcox, and J. Fristrom. 1993. The distribution of PS integrins, laminin A and F-actin during key stages in *Drosophila* wing development. *Development*. 117:509–523.
- George, E.L., E.N. Georges-Labouesse, R.S. Patel-King, H. Rayburn, and R.O. Hynes. 1993. Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. *Development*. 119:1079–1091.
- Goldstein, S.B., S. Gunawardena. 2000. Flying through the *Drosophila* cytoskeletal genome. *J. Cell Biol.* 150:F63–F68.
- Gotwals, P.J., L.I. Fessler, M. Wehrli, and R.O. Hynes. 1994a. *Drosophila* PS1 integrin is a laminin receptor and differs in ligand specificity from PS2. *Proc. Natl. Acad. Sci. USA.* 91:11447–11451.
- Gotwals, P.J., S. Paine-Saunders, K. Stark, and R.O. Hynes. 1994b. *Drosophila* integrins and their ligands. *Curr. Opin. Cell. Biol.* 6:734–739.
- Grotewiel, M.S., C.D.O. Beck, K.-H. Wu, X.-R. Zhu, and R.L. Davis. 1998. Integrin-mediated short-term memory in *Drosophila*. *Nature*. 391:455–460.
- Hutter, H., B.E. Vogel, J.D. Plenefisch, C.R. Norris, R.B. Proenca, J. Spieth, C. Guo, S. Mastwal, X. Zhu, J. Scheel, and E.M. Hedgecock. 2000. Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science*. 287:989–994.
- Hynes, R.O. 1992. Integrins: versatility, modulation and signalling in cell adhesion. *Cell*. 69:11–25.
- Hynes, R.O. 1996. Targeted mutations in cell adhesion genes: what have we learned from them? *Devel. Biol.* 180:402–412.
- Kreis, T., and R. Vale, editors. 1999. Guidebook to the Extracellular Matrix and Adhesion Proteins. Sambrook and Tooze Publishers, Oxford University Press, Oxford, UK.
- Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischman, et al. 2000. Comparative genomics of the eukaryotes. *Science*. 287:2204–2215.
- Samakovlis, C., G. Manning, P. Steneberg, N. Hacohen, R. Cantera, and M.A. Krasnow. 1996. Genetic control of epithelial tube fusion during *Drosophila* tracheal development. *Development*. 122:3531–3536.
- Stark, K.A., G.H. Yee, C.E. Roote, E.L. Williams, S. Zusman, and R.O. Hynes. 1997. A novel alpha integrin subunit associates with β PS and functions in tissue morphogenesis and movement during *Drosophila* development. *Development*. 124:4583–4594.
- Teppass, U. 1999. Genetic analysis of cadherin function in animal morphogenesis. *Curr. Opin. Cell Biol.* 11:540–548.
- Wheeler, G.N., and R.O. Hynes. 2000. The cloning, genomic organization and expression of the focal contact protein, paxillin, in *Drosophila*. In press.
- Yagi, T., and M. Takeichi. 2000. Cadherin superfamily genes: functions, genomic organization and neurologic diversity. *Genes Devel.* In press.
- Zavortink, M., T.A. Bunch, and D.L. Brower. 1993. Functional properties of alternatively spliced forms of the *Drosophila* PS2 integrin α subunit. *Cell Adhes. Commun.* 1:251–264.