

Shmuel Pietrokovski · Ben-Zion Shilo

## Identification of new signaling components in the *Drosophila* genome sequence

Received: 24 July 2000 / Accepted: 30 July 2000 / Published online: 14 September 2000  
© Springer-Verlag 2000

**Abstract** The availability of the complete sequence of the *Drosophila* genome and the assignment of putative reading frames, provides an opportunity to search for new members in families of proteins generating signaling cascades. The six major pathways that dictate patterning were examined: receptor tyrosine kinases, transforming growth factor beta (TGF $\beta$ ), Wnt, Toll, Hedgehog and Notch. Several new components were identified for the first four pathways, including ligands, receptors, cytoplasmic components and transcription factors. Most notable is the identification of a vascular endothelial growth factor (VEGF) receptor tyrosine kinase, two insulin/insulin growth factor I (IGF I) receptors without cytoplasmic protein kinase domains, and a family of proteins similar to Rho1 (a protein involved in cleavage of TGF $\alpha$ -like ligands). A new TGF $\beta$  family ligand, two new Wnts and a Frizzled receptor were also identified. Finally, for the Toll pathway, two new potential Spatzle-like ligands and two new receptors were identified.

**Keywords** *Drosophila* · Signaling · Receptors · Ligands

### Introduction

Embryonic and post-embryonic development in all multicellular organisms relies on elaborate communication between cells, which dictates the fates cells will adopt. These communication events are mediated by diffusible or membrane-bound ligands, triggering signaling cascades in the receiving cells. Saturating genetic screens for mutations affecting patterning identified most of the genes involved in these processes. Individual genes were then clustered into pathways, based on their structure, phenotype and epistasis studies. Elucidation of the com-

ponents of these cascades was boosted by parallel studies in *Drosophila* and *Caenorhabditis elegans*, as well as by the analysis of the vertebrate genes that were identified primarily through their role in oncogenic transformation. For each of the pathways uncovered, a comprehensive network of proteins was identified: proteins involved in ligand processing or presentation, ligands, transmembrane receptors, accessory proteins, and the cascade of proteins relaying the signal from the activated receptor to the nucleus, culminating in the induction of specific target genes. Surprisingly, the entire array of developmental decisions is regulated by only six signaling cascades that are conserved from flies to humans. While the molecular strategy varies dramatically between the different pathways, they all fulfill a similar basic function: transmitting the information from the membrane to the nucleus, and activating the appropriate sets of target genes.

The six cardinal pathways that play a role in development will be briefly mentioned below. Receptor tyrosine kinases (RTKs) are dimerized after binding to their ligands. This leads to transphosphorylation of cytoplasmic tyrosine residues, and the recruitment of the GTP exchange factor Sos to the membrane, triggering activation of the Ras protein (reviewed in Pawson 1995). Activated Ras triggers a cascade of three kinases: Raf, MEK and MAP kinase (MAPK). MAPK is translocated to the nucleus where it activates transcription by phosphorylating transcription factors.

TGF $\beta$  family proteins activate their pathway by binding to a type I receptor, and inducing its dimerization with a type II receptor. The cytoplasmic kinase of type II receptor phosphorylates the GS domain on the type I receptor. This allows the type I receptor to associate with a Smad protein and phosphorylate it. Once phosphorylated, Smad proteins heterodimerize with a co-Smad protein, translocate to the nucleus and participate in the transcription complex (reviewed in Massague and Chen 2000).

Wnt proteins activate seven-pass transmembrane receptors of the Frizzled family. Activated receptors affect

S. Pietrokovski · B.-Z. Shilo (✉)  
Department of Molecular Genetics,  
Weizmann Institute of Science, Rehovot 76100, Israel  
e-mail: Benny.Shilo@Weizmann.ac.il  
Tel.: +972-8-9343169, Fax: +972-8-9344108

the conformation of the Disheveled protein, which becomes capable of destabilizing a complex of proteins that normally target the Armadillo protein for degradation. The complex contains Axin, ZW3 kinase and APC. When this complex is destabilized, Armadillo levels increase allowing it to enter the nucleus and participate in a transcription complex with TCF (reviewed in Wodarz and Nusse 1998).

Hedgehog signals by binding the Patched 12-pass transmembrane receptor, thus dissociating Patched from the seven-pass transmembrane protein Smoothed. Dissociated Smoothed is active, and leads eventually to the proper cleavage and activation of the Ci protein which is the critical transcription factor (reviewed in Ingham 1998; Johnson and Scott 1998).

The Notch pathway is activated by binding of the ligands Delta or Serrate, facilitating cleavage of the intracellular domain of Notch. This domain is complexed with the Su(H) protein, and translocates to the nucleus where it activates transcription of target genes (reviewed in Artavanis-Tsakonas et al. 1999).

Finally the Toll receptor is activated by the ligand Spatzle (Spz). An elaborate cascade of serine proteases is responsible for generating the cleaved, active form of Spz. Activated Toll leads to the degradation of the I $\kappa$ B homologue, Cactus. This releases the NF $\kappa$ B homologue Dorsal from the cytoplasm, allowing it to translocate to the nucleus and activate target genes (reviewed in Belvin and Anderson 1996; Imler and Hoffmann 2000).

Completion of the *Drosophila* genome sequence provides an opportunity to revisit these cardinal signaling pathways (reviewed in Rubin et al. 2000). For each pathway we already know most of the components involved in transmitting the information. How many additional components of the known classes, which were not previously identified by genetic screens, can now be found in the genome? We will show that for some of the pathways, new components can be identified by analyzing the sequences. However, the number of new components is not large, demonstrating that the genetic screens were extremely effective in identification of the elements for these signaling pathways.

It should be stressed that this is only a first-pass analysis, which is based on the currently available gene predictions for the *Drosophila* genome. Problems that might exist include fully unpredicted genes, missing exons from predicted genes, and fusion of separate genes. Our analysis relies on sequence similarity between predicted *Drosophila* genes and known signaling pathway proteins. New types of signaling elements and those beyond the sensitivity of current sequence analysis methods would not be identified. Additional signaling proteins may be identified in the future as gene prediction and similarity search algorithms are refined, more EST sequence data are available, and gaps in the published data are sequenced. However, we believe that this initial analysis already provides a good estimate for the number of new genes that can be expected for each of the signaling cascades.

## Sequence analysis methods

The primary method we used was the BLAST sequence-to-sequence search method (<http://www.ncbi.nlm.nih.gov/blast>; Altschul et al. 1997). For a more sensitive analysis, the Blocks and InterPro database facilities for sequence-to-multiple-alignments and multiple-alignments-to-sequence searches were utilized (<http://www.ebi.ac.uk/interpro>; Apweiler et al. 2000; <http://blocks.fhcrc.org>; Henikoff et al. 2000). In order to identify genes we used the sequence analysis tools available on the Berkeley *Drosophila* Genome Project (BDGP) website ([http://www.fruitfly.org/seq\\_tools](http://www.fruitfly.org/seq_tools)). Transmembrane and signal-peptide regions were identified on the Center for Biological Sequence Analysis prediction servers (<http://www.cbs.dtu.dk/services>; Nielsen et al. 1997; Sonnhammer et al. 1998).

The data set we examined were the 14,100 *Drosophila melanogaster* protein sequences found in the NCBI Entrez database (<http://www3.ncbi.nlm.nih.gov/Entrez>) on 21 March 2000. When necessary we also examined the complete *D. melanogaster* nucleotide sequences found on the NCBI and BDGP databases.

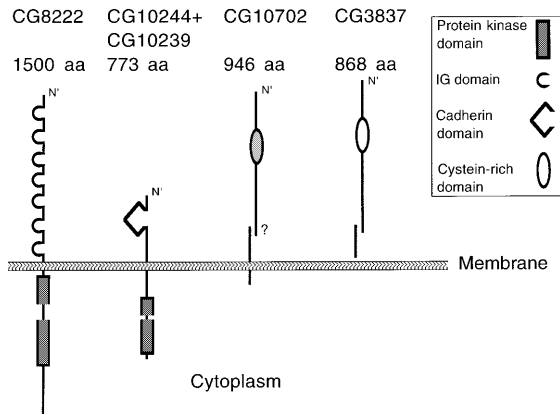
## RTKs

Newly identified proteins are listed in Table 1. Two new receptors were identified. One is a receptor which shows the highest similarity to vascular endothelial growth factor (VEGF) receptors in both ligand binding and protein kinase domains, across its 1,500 amino acids (accession no. CG8222) (Fig. 1). This may indicate an activity of a ligand with structural similarity to VEGF. Another new receptor (whose coding region is actually formed by the separately annotated CG10244 and CG10239 tandem ORFs) shows the greatest homology to the kinase domain of fibroblast growth factor (FGF) receptors in its intracellular domain. The extracellular domain of this receptor (designated in the annotation as another gene) contains a unique cadherin domain (Fig. 1). The only other known RTK family with this domain organization is the Ret proto-oncogene. However, no distinct similarity was found between the extracellular cadherin domain of Ret and this receptor.

An unexpected finding was the identification of receptors apparently lacking the intracellular domain. Two such receptors of the insulin/IGF I receptor family were identified. CG10702 contains a signal peptide and a transmembrane domain that is only 50 amino acids from the putative C-terminal end. No transmembrane region could be identified in the second protein (CG3837) (Fig. 1). In both cases no similarity to tyrosine kinase domains could be identified in the downstream nucleotide genomic regions. The significant similarity to the extracellular domain of insulin/IGF I receptors strongly suggests that these proteins are capable of binding ligands from the insulin/IGF I family. They may inhibit

**Table 1** Newly identified signaling-pathway proteins

Protein	Similar to	Comments
Receptor tyrosine kinases		
CG8222 CG10244 and CG10239	VEGF receptor Intracellular C' part similar to FGF receptor kinase domain and extracellular N' part similar to cadherin domains	See Fig. 1. Annotated as two separate ORFs (Fig. 1)
CG10702	Insulin/IGF I receptors	Signal peptide present, transmembrane region identified 50 amino acids from the putative C' end (Fig. 1)
CG3837	Insulin/IGF I receptors	Signal peptide present, no transmembrane region identified (Fig. 1)
CG6736, CG14167, CG8167, CG14173, CG14049, CG13317	Insulin-like proteins	First four located together at 67C1 and last two on chromosome X at 2F4 and 3F2, respectively
CG8056	Spitz, EGF-receptor ligand	–
CG3126, CG5522, CG8865, CG9491, CG3427, CG7369, CG4853	Sos-like guanosine exchange factors (GEFs)	–
CG5960	GTPase-activating protein (GAP)	–
CG12083, CG1214, CG1697, CG17212	Rhomboid	First two and Rhomboid are paralogues (see Fig. 2)
CG6892	ETS-domain protein	–
TGF $\beta$		
CG1901	TGF $\beta$ ligand	–
CG12410, CG11582	Twisted gastrulation (Tsg)	–
Wnt		
CG4971, CG4969	Wnt	–
CG4626	Frizzled receptor	–
CG17484	Armadillo	–
CG15010	Slimb	–
Hedgehog		
None	–	–
Notch		
None	–	–
Toll		
CG7250, CG8595	Toll receptor	–
CG18318, CG9972	Spatzle (Spz) ligand	–
CG1102	Easter protease	–
CG4394	Traf protein	–



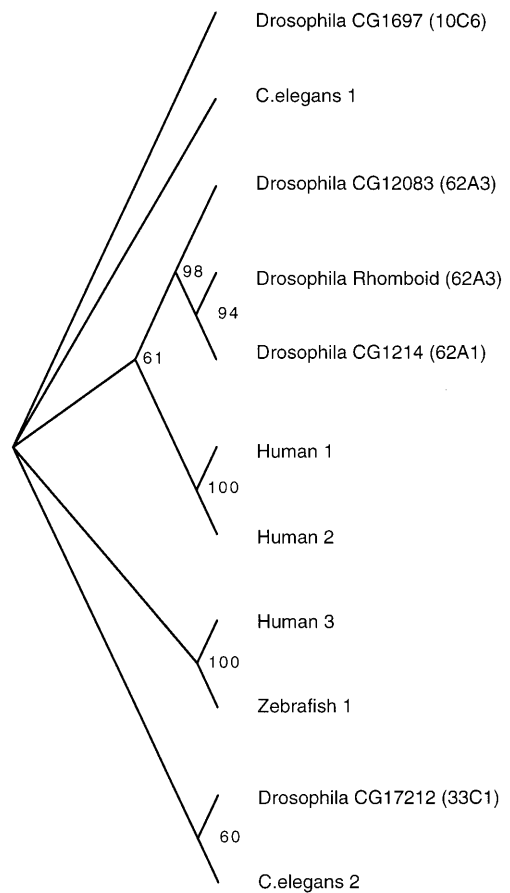
**Fig. 1** Schematic representation of newly identified RTK receptors. The discontinuity in CG10702 and CG3837 corresponds to their possible cleavage into insulin-receptors type  $\alpha$  and  $\beta$  subunits, however the cleavage site is less certain in CG10702

signaling by sequestering ligands. Alternatively they might promote signaling, for example by transporting or presenting the ligand to its receptor. The similarity between these two proteins (approximately 25% identity) is not greater than the similarity between the extracellular domains of various insulin receptors. Thus these proteins do not seem to be paralogues but to have evolved independently.

Several new RTK ligands were found. A new ligand of the EGF receptor termed Keren, displaying homology to Spitz, was identified (CG8056). Six insulin-like proteins were identified (CG6736, CG14167, CG8167, CG14173, CG14049, CG13317). The first four appear in tandem on chromosome 3L and the last two are on the X chromosome. While their genomic positions indicate gene duplication, the sequences of these proteins are diverse from each other and other insulin-like proteins, including insect prothoracicotrophic hormones (bombyxins) and the locust insulin-related peptide (LIRP). The ligand of the Heartless FGF receptor (Beiman et al. 1996; Gisselbrecht et al. 1996) which is expected to belong to the FGF family, was not identified, even by advanced blocks searches.

Several new potential downstream components were identified, including seven Sos homologues (CG3126, CG5522, CG8865, CG9491, CG3427, CG7369, CG4853) which are likely to be guanine exchange factors (GEFs) working either with Ras or other small GTP-binding proteins, and several Ras family homologues. Finally, another GTPase-activating protein (GAP) (CG5960) was identified.

It is interesting to highlight the discovery of a new gene family of Rhomboid (Rho) proteins. Rho is a seven-transmembrane protein (Bier et al. 1990), which is required for cleavage of the EGF receptor ligand Spitz (Schweitzer et al. 1995; Golembo et al. 1996; Bang and Kintner 2000). Four additional Rho homologues were identified (CG12083, CG1697, CG1214, CG17212). Database searches found Rho homologues in other spe-



**Fig. 2** Sequence relations among different Rhomboid-like proteins. Common sequence segments were identified and aligned by the clustalW program (Thompson et al. 1994), which was used to compute the tree and its significance (bootstrap) values. Branch points with less than 60% bootstrap values (out of 1,000 trials) were collapsed. The remaining branch points are each labeled by their *percentage bootstrap values*. Proteins were named by species name and an arbitrarily ordered number. The chromosomal locations of the *Drosophila* genes are shown in parentheses. NCBI gene identifier (*gi*) accession numbers of the non-*Drosophila* sequences: *Caenorhabditis.elegans* 1–529201; *C.elegans* 2–465822; Zebrafish 1–6950276; Human 1–4506525; Human 2–3935221 (exons of the apparent gene on the 5' region of this nucleotide sequence); Human 3–7020534

cies as well. While *Drosophila* has at least five Rho homologues (Wasserman et al. 2000) *C. elegans* includes only two and human at least three. We identified common conserved sequence regions present in all Rho sequences and used them to determine the relation between the proteins. Three of the *Drosophila* Rhos are clearly most similar to each other and are probably paralogues resulting from gene duplication following the separation of insects, vertebrates and nematodes. These three proteins might be related to two human Rhos which are paralogues of each other (Fig. 2). These two independent duplication events, apparent in *Drosophila* and humans may have resulted from similar modes of action or selection of Rho type proteins.

Pointed is an ETS-domain protein which is a cardinal player in the downstream responses to the EGF receptor

and Sevenless signaling. Five other ETS proteins are known and an additional one (CG6892) was identified, showing similarity in the ETS DNA binding domain.

---

### TGF $\beta$ signaling

Several TGF $\beta$  family ligands were already known: Dpp, Screw, Glass bottom boat, myoglianin and  $\beta$ -activin precursor. One additional protein (CG1901) has been identified, which shows the similarity in its C-terminal 230 amino acids, including the domain which binds the receptor after its proteolytic processing. The cysteine residues which are essential for the structure have been conserved.

Twisted gastrulation (Tsg) is involved in the facilitation and regulation of Short gastrulation cleavage by Tolloid (Oelgeschlager et al. 2000; Yu et al. 2000). Two new Tsg homologues have been identified (CG12410, CG11582), showing a conserved pattern of more than 20 cysteines.

---

### Wnt pathway

Several Wnt ligands are known. Two new Wnt homologues (CG4971, CG4969) were found. A new Frizzled (Frz) receptor was identified (CG4626). It has the typical structure of an N-terminal cysteine-rich domain and seven transmembrane domains. This additional receptor may be responsible for interaction with the new Wnt members, or may function cooperatively with Frz1 or Frz2 in the specification of cell fates and establishment of planar polarity.

Regulation of Armadillo (Arm) stability is at the heart of the Wnt pathway. A new Arm homologue was identified (CG17484), which contains Arm repeats and is most similar to neural plakoglobin-related protein. This protein may participate in Wnt signaling or in the establishment of the adherence junctions.

Slimb protein provides specificity of Arm recognition by the ubiquitination/degradation machinery (Jiang and Struhl 1998). A homologue of Slimb was identified (CG15010), showing similarity along the entire length of Slimb.

---

### Toll pathway

The Toll pathway was initially dissected on the basis of its role in specification of the embryonic dorso-ventral axis (Belvin and Anderson 1996). Unraveling the components has revealed a striking similarity to the pathway activated by IL1 receptor, leading to the nuclear translocation of NF $\kappa$ B, involved in the immune response. Subsequently, the involvement of the Toll pathway in the *Drosophila* immune response has been demonstrated. Finally, vertebrate Toll-like receptors contributing to the innate immune response were identified (Hoffmann et al.

1999). Among the signaling pathways we examined, this is the one in which the largest number of new components have been identified. These elements may represent hitherto uncharacterized components of the immune response in *Drosophila*.

Spz is the ligand activating Toll in the ventral side of the embryo and is also involved in the immune response. Two new Spz homologues were found (CG18318, CG9972). Interestingly, they display the similarity in the C-terminal part, which is the active ligand domain of Spz after its cleavage by Easter. These are likely to be activated by proteolytic cleavage at the onset of the immune response. Several known Toll homologues, 18 wheeler, Tahoe and Tak1 are known. Two additional homologues were identified (CG7250, CG8595). They are similar in length, and show homology both in the extracellular leucine-rich region as well as in the cytoplasmic domain.

A proteolytic cascade is responsible for activating the cleavage of Spz. Among the large number of new serine proteases identified in the genome (Rubin et al. 2000), one shows a high level of similarity to Easter along the entire length (CG1102). Again, this protease may be involved in the immune response, as Easter was shown not to be involved in this process. Traf is an essential adaptor protein relaying the signals from Toll receptors. In addition to the two known Traf proteins, a new member was found (CG4394).

### Concluding remarks

A search of the completed genome sequence of *Drosophila* has identified new homologues of signaling components in several cardinal signaling pathways. It was surprising however, that the number of new components is fairly small. For the Hedgehog and Notch pathways no new components were identified.

Most of the known components of the signaling pathways were previously identified by genetic screens, including direct identification of mutant embryonic or post embryonic phenotypes, and sensitized genetic screens. The key to identification of such phenotypes is a non-redundant role for each of the respective genes. The number of new components that may be identified by data mining of the complete genome sequence is thus an indication of the degree of functional redundancy. In addition, it may identify essential components in biological processes that were not searched exhaustively for mutants, such as the immune response.

The relatively small number of new components identified in *Drosophila* strongly suggests that the degree of functional redundancy in the components of the major signaling pathways is minimal.

**Acknowledgments** This work was supported by grants from a Kekst Family Center for Medical Genetics research grant, the Helen and Milton Kimmelman Center, the Forchheimer center and the Crown Human Genome Center of the Weizmann Institute to S.P., and from the German-Israeli Foundation and the US-Israel Binational Foundation to B.S.

## References

- Altschul SF, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Apweiler R, et al (2000) InterPro - an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* (in press)
- Artavanis-Tsakonas S, et al (1999) Notch signaling: cell fate control and signal integration in development. *Science* 284:770–776
- Bang AG, Kintner C (2000) Rhomboid and Star facilitate presentation and processing of the *Drosophila* TGF- $\alpha$  homologue Spitz. *Genes Dev* 14:177–186
- Beiman M, et al (1996) Heartless, a *Drosophila* FGF receptor homolog, is essential for cell migration and establishment of several mesodermal lineages. *Genes Dev* 10:2993–3002
- Belvin MP, Anderson KV (1996) A conserved signaling pathway: the *Drosophila* toll-dorsal pathway. *Annu Rev Cell Dev Biol* 12:393–416
- Bier E, et al (1990) *rhomboid*, a gene required for dorsoventral axis establishment and peripheral nervous system development in *Drosophila melanogaster*. *Genes Dev* 4:190–203
- Gisselbrecht S, et al (1996) *heartless* encodes a fibroblast growth factor receptor (DFR1/DFGF-R2) involved in the directional migration of early mesodermal cells in the *Drosophila* embryo. *Genes Dev* 10:3003–3017
- Golembo M, et al (1996) The *Drosophila* embryonic midline is the site of Spitz processing, and induces activation of the EGF receptor in the ventral ectoderm. *Development* 122:3363–3370
- Henikoff JG, et al (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* 28:228–230
- Hoffmann JA, et al (1999) Phylogenetic perspectives in innate immunity. *Science* 284:1313–1318
- Imler JL, Hoffmann JA (2000) Signaling mechanisms in the antimicrobial host defense of *Drosophila*. *Curr Opin Microbiol* 3:16–22
- Ingham PW (1998) Transducing Hedgehog: the story so far. *Embo J* 17:3505–3511
- Jiang J, Struhl G (1998) Regulation of the Hedgehog and Wingless signaling pathways by the F-box/WD40-repeat protein Slimb. *Nature* 391:493–496
- Johnson RL, Scott MP (1998) New players and puzzles in the Hedgehog signaling pathway. *Curr Opin Genet Dev* 8:450–456
- Massague J, Chen YG (2000) Controlling TGF- $\beta$  signaling. *Genes Dev* 14:627–644
- Nielsen H, et al (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10:1–6
- Oelgeschlager M, et al (2000) The evolutionarily conserved BMP-binding protein Twisted gastrulation promotes BMP signaling. *Nature* 405:757–763
- Pawson T (1995) Protein modules and signaling networks. *Nature* 373:573–580
- Rubin GM, et al (2000) Comparative genomics of the eukaryotes. *Science* 287:2204–2215
- Schweitzer R, et al (1995) Secreted Spitz triggers the DER signaling pathway and is a limiting component in embryonic ventral ectoderm determination. *Genes Dev* 9:1518–1529
- Sonnhammer E, et al (1998) In: Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C (eds) A hidden Markov model for predicting transmembrane helices in protein sequences. Sixth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, Calif, pp. 175–182.
- Thompson JD, et al (1994) CLUSTAL W:improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Wasserman JD, et al (2000) A family of *rhomboid*-like genes:*Drosophila rhomboid-1* and *roughoid/rhomboid-3* cooperate to activate EGF receptor signaling. *Genes Dev* 14:1651–1663
- Wodarz A, Nusse R (1998) Mechanisms of Wnt signaling in development. *Annu Rev Cell Dev Biol* 14:59–88
- Yu K, et al (2000) Processing of the *Drosophila sog* protein creates a novel BMP inhibitory activity. *Development* 127:2143–2154