Approaching *Drosophila* development through proteomic tools and databases: at the hub of the post-genomic era

Ana Carmena

Instituto de Neurociencias de Alicante/CSIC-UMH, Sant Joan d´Alacant, 03550 Alicante, Spain.

Key words: Proteomics, *Drosophila*, development, databases, bioinformatics tools, interactome networks

Corresponding author: acarmena@umh.es

Phone: +34-965919230

Fax: +34-965919561

Abstract

The past decade has witnessed an explosion in the growth of proteomics. The completion of numerous genome sequences, the development of powerful protein analytical technologies, as well as the design of innovative bioinformatics tools have marked the beginning of a new post-genomic era. Proteomics, the large-scale analysis of proteins in an organism, organ or organelle encompasses different aspects: 1) the identification, analysis of post-translational modifications and quantification of proteins; 2) the study of protein-protein interactions; and 3) the functional analysis of interactome networks. Here, we briefly summarize the emerging analytical tools and databases that are paving the way for studying *Drosophila* development by proteomic approaches.

1. Introduction

Drosophila melanogaster has proved to be an excellent model organism to investigate the fundamental principles of development, from classical genetics and embryology analyses to molecular biology approaches. As a well-established model organism, the *Drosophila* genome was one of the first to be sequenced (Adams et al., 2000; Myers et al., 2000) and since then, different large-scale studies have helped greatly to improve the annotation of the fly genome (Celniker et al., 2002; Hoskins et al., 2002; Lin et al., 2007; Stapleton et al., 2002a; Stapleton et al., 2002b). In the post-genomic era, proteomics has become central to systems biology although its implementation still remains a continuous challenge. Proteomic approaches involve the large-scale identification and quantification of proteins, the study of their post-translational modifications, the analysis of protein-protein interactions (PPIs) as well as *in vivo* studies of protein expression in order to identify functional protein networks. An ever-increasing number of bioinformatics tools are being developed to achieve these goals. These powerful techniques will facilitate the complete mapping of proteomes and the identification of biologically significant interactome networks. To fulfill this last aim, it is of great relevance to combine and to integrate information from different datasets. Nowadays, we can start to achieve this task. For example, only some years ago the first comprehensive PPI map was completed in *Drosophila* by a genome-wide yeast two-hybrid approach (Giot et al.,2003). Later on, proteome-wide purification of protein complexes was accomplished in *Drosophila* with tandem affinity purification mass spectrometry (TAP-MS) (Veraksa et al., 2005). Regardless some disadvantages and difficulties, these are still very efficient approaches, which are continuingly improving, to unveil PPI maps (Kyriakakis et al., 2008). However, although these proteome-wide analyses provide first predictions for protein function, they do not

yield, for example, any information about spatio-temporal protein expression. The recent developing of *Drosophila* databases that uncover protein expression patterns can help to fill that gap and thus to yield more accurate predictions of functional PPIs (Tomancak et al., 2007). Moreover, the inclusion in the databases of Gene Ontology (GO) annotations and information from different organisms datasets can help to further filter biologically significative PPIs (see below). Proteomic studies will also deepen our knowledge of genomes, improving their annotation and our understanding of genomic regulatory networks and their evolution. Here, we review the latest bioinformatics tools and datasets that are being applied to the large-scale proteomic analysis of *Drosophila* development.

2. Identification and quantification of proteins: The *Drosophila* PeptideAtlas

The characterization and quantitative measurement of whole proteomes is a critical yet difficult issue of proteomic approaches. Indeed, no proteome has yet been fully mapped. A fly proteomics and genomics resource, the *Drosophila* PeptideAtlas was recently developed to help to fulfill this aim (Brunner et al., 2007; Loevenich et al., 2009). PeptideAtlas is an open-access database (http://www.drosophila-peptideatlas.org) and hitherto the largest fly proteome catalogue described. It contains about 76,724 peptide sequences representing 9,263 protein isoforms (compared to the 1,552 fly protein sequences listed in SwissProt) and there are 8,799 gene models, which is the equivalent to 65% of the Ensembl database (v27.3c). This gene model coverage is 13% higher than that of the human Peptide Atlas, which includes 52% of the Ensembl V43 genes.

A "two-step strategy" has been followed to quantitatively establish the proteome catalogue (Ahrens et al., 2007; Brunner et al., 2007). In a first step, shotgun

proteomic experiments provided the mass spectrometry identification of peptides from complex samples. In a second step, a peptide catalogue was established (and will be continuously expanded by introducing information from other databases and sources). With all this information, a minimal set of peptides that represent each protein in the catalogue was selected. These peptides are called ProteoTypic Peptides (PTPs). A consensus mass spectrum was calculated for each PTP to represent the proteome catalogue avoiding redundancy. Through a targeted mass spectrometry method called multiple Selected Reaction Monitoring (mSRM), PTPs can then be identified and quantified in complex mixtures (Ahrens et al., 2007; Anderson and Hunter, 2006; Kuster et al., 2005). The fly proteome catalogue is not yet complete. As mentioned above, proteins corresponding to 35% of the predicted gene models remain uncovered, as some limitations still exist and full proteome coverage is difficult to achieve (Brunner et al., 2007; Castellana et al., 2008). Nevertheless, as more protein extracts are analyzed in the future, the catalogue will be more comprehensive. By providing protein specific peptides and their representative consensus mass spectra, the proteome catalogue will enormously facilitate proteomics experiments in *Drosophila* (Fig. 1). Furthermore, PeptideAtlas information can be used to revisit gene models and to improve genome annotation.

3. Post-translational protein modifications: PhosphoPep

An important coverage of proteomics is the analysis of post-translational protein modifications (PTMs). These modifications are key for multiple cellular processes during development. For example, protein phosphorylation/dephosphorylation plays a central role in many signaling networks and it is a process precisely regulated by kinases and phosphatases. In *Drosophila*, a phosphoproteome resource called

PhosphoPep has been recently developed (Bodenmiller et al., 2007). This database contains more than 10,000 phosphorylation sites mapping to 4,583 phosphoproteins (3,472 gene models) of *Drosophila* Kc167 cell line. PhosphoPep (http://www.phosphopep.org) can be searched through several interfaces that provide different type of information about a protein of interest. For example, for each protein it is shown the sequence of the observed phosphopeptides with the phosphorylation site(s). Additional data about each phosphopeptide such as the probability (PeptideProphet probability score), mass, how many times was observed and how many proteins it maps can also be found. The phosphopeptides are highlighted within the protein sequence where the ambiguous phosphorylation sites are represented with different colour. PhosphoPep is to date the most comprehensive database of this kind. Very recently, Zhai and colleagues have also analyzed the phosphoproteome of *Drosophila* but using embryos instead of Kc167 cells as a sample (Zhai et al., 2008). Intriguingly, the overlap between this and the previously mentioned PhosphoPep dataset was very low. For example, whereas PhosphoPep identified 4,583 phosphoproteins and found 13% of peptides phosphorylated at multiple sites, the study of Zhai and colleagues only describes 2,702 phosphoproteins and the fraction of multiply phosphorylated peptides detected was 68%. Different samples (i.e. cells versus embryos), methods and analysis tools can explain the differences found in both studies. Hence, it will be of great relevance for the future to take into account the combination of different approaches to unveil the complexity of phosphoproteomes in particular and of proteomics in general.

4. Protein-protein interaction (PPI) networks

Identifying which are the interacting partners of a given protein is crucial. These partners and the resulting interactome network are fundamental to understand cell behavior. Proteomics is contributing significantly to define the protein networks through large-scale interaction studies, such as high-throughput yeast two-hybrid screens or affinity-based protein complex purification. Taking advantage of all the information generated, different databases and models have been designed to predict interactome networks in *Drosophila melanogaster.*

4.1. Fly-DPI: Database of Protein Interactomes

Fly-DPI is an integrated proteomic tool that combines statistical and biological estimates to validate PPIs (http://flydpi.nhri.org.tw) (Lin et al., 2006). The statistical reference system is a hybrid of the Maximum Likelihood Estimation (MLE) and association methods. This hybrid model predicts PPIs based on experimental data about the interactions between different protein domains. For example, the model considers that two proteins interact if at least one pair of their domains can associate, providing an estimated probability of the interaction for each domain pair. Comprehensive biological annotations in major databases enable the establishment of putative functional connections between known annotated proteins and their predicted partners with unknown function. Spatio-temporal information about proteins can also be used as a "biological filter" to reinforce the probability of an interaction. A novel tool provided by the Fly-DPI database is the "ping-pong" search. This search identifies the shortest path between any two proteins, showing the networks associated with both. Thus, Fly-DPI as a whole can provide an estimate of PPIs based on both experimental and predicted data. Biological annotations then contribute to increase the reliability of any given PPI.

4.2. DroID: the Drosophila Interactions Database

DroID is probably the most comprehensive interactions database available for *Drosophila* (http://www.droidb.org) (Yu et al., 2008). This is a highly extended version of an earlier database (Pacifico et al., 2006) that will continue to be periodically updated. The predicted PPIs are derived, in part, from experimental data from other model organisms and humans. These predicted interactions are called "interologs" (Yu et al., 2004b) and since the conservation of PPIs in different species is common, it represents potentially useful information. In addition to computationally predicted PPIs, the DroID database takes into account reported physical PPIs, genetic interactions and gene expression data. A confidence score between 0 and 1 is assigned to each interaction. The higher the score, the higher is the probability that the interaction is biologically relevant (Fig. 2). This confidence score is updateable and represents a novelty of DroID to annotate each physical PPI (more information about the method followed to assign the confidence score can be found in Giot et al., 2003). Gene expression and expression correlation data are also included in DroID. These data can be useful to search and filter interactions, as well as to define specific gene sub-networks (Arbeitman et al., 2002; Tomancak et al., 2002). Genes that are co-expressed are more likely to function in the same biological process and thus, in DroID, each gene pair is annotated with gene expression correlation values. Indeed, proteins that physically interact are encoded by gene pairs with higher expression correlation values that random gene pairs. Such correlations have been also described previously in *S.cerevisiae* (Ge et al., 2001). Gene pairs that interact genetically, another parameter that DroID exploits, also show higher expression correlation values and so, they have more probability to work together in a given process. In addition, all the original experiments and sources are present in the

annotated interactions. DroID is an example of the need and utility of comprehensive databases focused on particular model organisms to analyze biology systems. Other databases are emerging that attempt to offer such comprehensive coverage for other specific organisms, such as HomoMINT and UniHI for humans (Chaurasia et al., 2007; Persico et al., 2005). Also, some interactome analysis methods developed previously (Baudot et al., 2006; Brun et al., 2003) have more recently been applied to decipher the *Drosophila* interactome. For example, this is the case of the graph-theory based method called PRODISTIN, a tool to functionally classify proteins within interactome networks (Baudot et al., 2008).

4.3 A hub protein classifier: a bioinformatics tool to predict highly-connected nodes within PPI networks

Most PPI networks are scale-free, which means that a large number of proteins are poorly connected and only a small number of proteins interact with many partners. These few, highly-connected nodes within PPI networks are called hubs (Albert, 2005; Barabasi and Oltvai, 2004). As a consequence of their non-homogeneous architecture, networks are very vulnerable to the targeted removal of hubs, although they are extremely robust to random failures (Albert et al., 2000). This "centrality-lethality" rule has been observed in PPI networks from different organisms, including yeast, nematodes and flies, and it reflects the relevance of the network structure and the topological positions of individual proteins (Albert et al., 2000; Hahn and Kern, 2005; Jeong et al., 2001; Wuchty, 2002; Yu et al., 2004a). An alternative explanation of the centrality-lethality rule has been proposed since not all PPIs are equally critical for the cell (He and Zhang, 2006). Indeed, only a small percent of interactions are essential. Authors of this work claim that hubs are few and highly-connected nodes

but randomly distributed in the network (i.e. independent of network architecture). They explain the centrality-lethality rule by arguing that hubs are important since by having more partners, the probability of being involved in essential PPIs is also higher (He and Zhang, 2006). Although the position of hubs within a network may not be considered crucial for some networks, in other biological networks the influence of network architecture cannot be disregarded. Hubs are, in any case, critical for the organization, function and survival of PPI networks (Albert et al., 2000; Jeong et al., 2001).

The hub classifier developed by Hsing and colleagues (Hsing et al., 2008) takes advantage of GO protein annotations and protein-protein interaction data. GO annotations can reflect important functional properties of proteins, even when the annotations are not based on experimental data (Ashburner et al., 2000; Camon et al., 2004; Qi et al., 2006; Rhee et al., 2008). Hence, the hub classifier could be also used as a predictive bioinformatics tool in organisms for which PPI data is not available. To develop this hub classifier, annotated proteins of *E.coli*, *S.cerevisiae*, *D.melanogaster* and *H.sapiens* were successfully used, showing that highly connected nodes share functional properties that appear in their GO annotations (e.g., translation, cell cycle, cell death or signal transduction). In addition, the hub classifier had the highest predictive value when compared with other protein prediction tools (Hsing et al., 2008). Given the particular characteristics of hub proteins, predicted hubs can be used for in depth analysis of cellular processes and to identify novel drug targets. Likewise, they can be used to select baits for large-scale experiments to pull-down protein complexes. The hub protein classifier can be searched at: http://www.cnbi2.ca/hub/.

5. Functional interactome networks: protein localization and spatio-temporal regulation

One critical means to decipher functional PPI networks during development is to analyze the spatio-temporal expression patterns of the potential partners. Databases and GO annotations can be extremely useful and accurate in making PPI predictions. However, co-localization of the predicted potential partners is a prerequisite to consider the interactions relevant *in vivo*. Indeed, information about protein expression is frequently used to filter datasets when predicting PPIs (see above).

In *Drosophila*, an embryonic gene expression atlas has been developed on a genome-wide basis (Tomancak et al., 2002). A recent release of this database includes about 6,003 (44%) of the 13,659 protein encoding *Drosophila* genes, a representative sample of the whole genome (Tomancak et al., 2007). This is the most comprehensive dataset of its kind to be developed for a multicellular organism. Two methods were employed in parallel to implement the dataset: RNA in situ hybridization and microarray analysis. In situ hybridization provides "qualitative", spatial information regarding gene expression during embryogenesis, whereas microarrays reveal quantitative temporal measures of gene expression. The results of both datasets are integrated and they complement one another. The expression patterns were classified into large clusters defined by controlled vocabulary (CV) annotations for embryonic anatomy (Grumbling and Strelets, 2006). To compare the characteristics of gene expression for a set of clustered genes, authors create a linear representation of the CV annotations called the "anatomical signature" or "anatogram". With this representation, one can visualize where and when a given gene set is expressed during development, and compare this with the spatio-temporal profile of other gene sets. Moreover, the expression patterns of different species can also be analyzed with

anatograms. All the expression patterns originally analyzed were grouped into two main categories: broad (representing more than half of the genes) and tissue-restricted. Many of the genes (41%) expressed in restricted tissues were classified in multiple clusters (e.g., Fas-3 belongs to three different clusters: "early epithelial pattern", "visceral muscle" and "midline/CNS cluster"). Other surveys have been carried out in *Drosophila* to study gene expression patterns in more detail, focusing on specific embryonic stages or on specific protein-encoded genes (Gurunathan et al., 2004; Keranen et al., 2006; Luengo Hendriks et al., 2006; Ye et al., 2006). The genome-wide dataset of *D.melanogaster* embryonic expression patterns (http://www.fruitfly.org/cgi-bin/ex/insitu.pl) will promote and facilitate the in depth analysis of gene and protein regulatory networks during development.

One inherent problem to these approaches is the huge amount of data generated. Very recently, a new web interface, the CATMAID (Collaborative Annotation Toolkit for Massive Amounts of Image Data) has been developed (Saalfeld et al., 2009). The CATMAID will be of great help to organize, annotate and to browse big datasets, including the large scale dataset of *Drosophila* embryonic expression patterns (Tomancak et al., 2007). In addition, researchers can share their image datasets (e.g. confocal, electron microscopy or in situ image data) uploading them into the CATMAID. This collaborative annotation can notably improve protein expression data annotation. Similar approaches to the genome-wide dataset of *D.melanogaster* embryonic expression patterns in other species can also impinge on our knowledge about evolutionary conserved expression patterns and functional interactome networks. Indeed, comparision and combination of different datasets within an organism and between different model organisms can be extremely helpful

for interpreting data and drawing conclusions. This consideration was the base for developing a new source for integrated data: FlyMine.

6. FlyMine: combining datasets

One challenge of system-wide approaches in general is how to integrate the huge amount of data generated. For that task, it is of great relevance combining information of different datasets both within the same organism as well as between different species. FlyMine (http://www.flymine.org/) was developed few years ago as such integrative database (Lyne et al., 2007). Although initially FlyMine was mainly focused on *Drosophila* and *Anopheles* genomic and proteomic datasets, it currently (release 18.0) includes information about other species as *C.elegans*, *S.cerevisiae*, *M.musculus* and *H.sapiens*. Combining and integrating data is a powerful mean to cross-validate datasets between different organisms. For example, predicted interologs (pairs of proteins whose orthologs in another specie interact) can increase confidence to interactions and also can generate new hypotheses to study. Some of the *Drosophila* databases that uncover PPIs mentioned before, such as Dro-ID, also include information from other organisms. FlyMine, in addition to PPIs, provides different data sources such as genome and proteome annotation, comparative genomics, protein structure, phenotypes, "Homophila" (Human disease to *Drosohila* database), among others (Fig. 3). New releases of FlyMine appear periodically with continuous improvements, additional data sources and updated information. FlyMine is linked to other databases, including the primary database for any drosophilist: FlyBase.

7. FlyBase: the hub of drosophilists databases

FlyBase (http://flybase.bio.indiana.edu/) is widely known and recognized by the *Drosophila* community as the main fly database it represents. It is important though to highlight the huge impact that FlyBase has had and is having for the creation and expansion of genome-wide and proteome-wide tools and so, for the large-scale analysis approach of *Drosophila* development. Indeed, Flybase stands as the central database to which the other *Drosophila* tools mentioned before, such as the Peptide-Atlas, Dro-ID, PhosphoPep or the FlyMine, are linked with. Fly-DPI database is the only case in which it is not shown a clear link with FlyBase, even though nomenclature and GO terms are taken from there. This should be revised to facilitate cross-talk between datasets.

8. Perspectives

All tools and databases presented here have limitations and are susceptible of improvements. A clear sign of this is the constant updating and the new releases of most databases that periodically appear. Proteomic techniques per se still have a number of limitations, such as resolution and sensitivity, albeit impressive advances have been accomplished in MS along past years. These techniques also confront important challenges, such as deeper proteome coverage. Indeed, today it has become apparent that different protocols and techniques reveal only a minimal part of the total proteome and, at present, no proteome has been completely characterized. However, the two-step strategy followed for the analysis of *Drosophila* proteome has been particularly successful, representing the first high-coverage proteome map for a multi-cellular eukaryote (Loevenich et al., 2009). Developing proteome catalogs for other organisms, similar to that created for *Drosophila* could help to increase the efficiency

of proteome characterization. Another aspect to take into account that we have already learned related to PPIs or PTMs coverage is how similar experiments can yield some overlapping but different set of results. Normally, this does not mean that one of the analyses is wrong but that different technique or approach has been selected. Hence, it is fundamental to combine several studies and datasets to try to eliminate false positives or compensate false negatives. Likewise, it is necessary to use biological filters such as GO annotations, protein expression patterns, interologs, etc, and combine all these parameters to establish good confidence scores for PPIs. Finally, the management of huge datasets we are tackling with today demands the optimization of data tools organization, integration and presentation. Hence, it is of great importance to develop user-friendly search pages for routine use. This will be an ongoing challenge in bioinformatics.

Large-scale datasets are important for researchers as an in-route to detailed analyses and validation of interactomes during development. Likewise, datasets can be used for modeling cell behaviors by in silico approaches. Obviously, cells are extremely complex, non-linear systems, and to be able to very precisely predict global cellular responses in multicellular organisms, after specific cell manipulations, is far away to be reached. Nevertheless, it is quite remarkable how biology already co-exists with mathematical and physical approaches to untangle cell complexity. This phenomenon has allowed, for example, the emergence and expansion of potent proteomic bioinformatics tools. In addition, physical and mathematical approaches are making possible the development of quantitative analyses to measure cell changes in particular processes in response to specific environment modifications. In a future, the intertwining of biology, mathematics and physic may succeed in making predictions about more complex aspects of cell behavior.

9. Conclusions

In a systems biology era, proteomics is emerging as a critical large-scale approach to analyze and understand complex developmental processes. The characterization and quantification of proteins, their interactions and their spatio-temporal expression patterns will provide invaluable information in this regard. The implementation of innovative bioinformatics tools and advanced technologies is contributing to reach this goal, as shown here for the well-established model organism *Drosophila melanogaster*. One important challenge of proteomics is to manage and to integrate large-scale data sets as well as to validate PPIs by alternative approaches, particularly those that can be applied in vivo. This will be crucial to uncover functionally relevant interactome networks.

Acknowledgements

I would like to thank anonymous reviewers for very helpful suggestions and comments on the manuscript. Work in my lab is supported by Grants from the Spanish Government BFU2006-09130 and CONSOLIDER-INGENIO 2010 CSD2007-00023.

Bibliography

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C. et al. (2000) The genome sequence of Drosophila melanogaster. Science 287, 2185-95.

Ahrens, C.H., Brunner, E., Hafen, E., Aebersold, R. and Basler, K. (2007) A proteome catalog of Drosophila melanogaster: an essential resource for targeted quantitative proteomics. Fly (Austin) 1, 182-6.

Albert, R. (2005) Scale-free networks in cell biology. J Cell Sci 118, 4947-57.

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. Nature 406, 378-82.

Anderson, L. and Hunter, C.L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 5, 573-88.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W. and White, K.P. (2002) Gene expression during the life cycle of Drosophila melanogaster. Science 297, 2270-5.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-9.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101-13.

Baudot, A., Angelelli, J.B., Guenoche, A., Jacq, B. and Brun, C. (2008) Defining a modular signalling network from the fly interactome. BMC Syst Biol 2, 45.

Baudot, A., Martin, D., Mouren, P., Chevenet, F., Guenoche, A., Jacq, B. and Brun, C. (2006) PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks. Bioinformatics 22, 248-50.

Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L.N., Shannon, P.T., Pedrioli, P.G., Panse, C., Lee, H.K., Schlapbach, R. and Aebersold, R. (2007) PhosphoPep--a phosphoproteome

resource for systems biology research in Drosophila Kc167 cells. Mol Syst Biol 3, 139.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A. and Jacq, B. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5, R6.

Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P.G., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A.J., Hafen, E., Schlapbach, R. and Aebersold, R. (2007) A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol 25, 576-83.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32, D262-6.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V. and Briggs, S.P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 105, 21034-8.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., Hodgson, A., George, R.A., Hoskins, R.A., Laverty, T., Muzny, D.M., Nelson, C.R., Pacleb, J.M., Park, S., Pfeiffer, B.D., Richards, S., Sodergren, E.J., Svirskas, R., Tabor, P.E., Wan, K., Stapleton, M., Sutton, G.G., Venter, C., Weinstock, G., Scherer, S.E., Myers, E.W., Gibbs, R.A. and Rubin, G.M. (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol 3, RESEARCH0079.

Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E.E. and Futschik, M.E. (2007) UniHI: an entry gate to the human protein interactome. Nucleic Acids Res 35, D590-4.

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29, 482-6.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., Jr., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M. (2003) A protein interaction map of Drosophila melanogaster. Science 302, 1727-36.

Grumbling, G. and Strelets, V. (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34, D484-8.

Gurunathan, R., Van Emden, B., Panchanathan, S. and Kumar, S. (2004) Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations. BMC Bioinformatics 5, 202.

Hahn, M.W. and Kern, A.D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22, 803-6.

He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? PLoS Genet 2, e88.

Hoskins, R.A., Smith, C.D., Carlson, J.W., Carvalho, A.B., Halpern, A., Kaminker, J.S., Kennedy, C., Mungall, C.J., Sullivan, B.A., Sutton, G.G., Yasuhara, J.C., Wakimoto, B.T., Myers, E.W., Celniker, S.E., Rubin, G.M. and Karpen, G.H. (2002) Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol 3, RESEARCH0085.

Hsing, M., Byler, K.G. and Cherkasov, A. (2008) The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. BMC Syst Biol 2, 80.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature 411, 41-2.

Keranen, S.V., Fowlkes, C.C., Luengo Hendriks, C.L., Sudar, D., Knowles, D.W., Malik, J. and Biggin, M.D. (2006) Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. Genome Biol 7, R124.

Kuster, B., Schirle, M., Mallick, P. and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 6, 577-83.

Kyriakakis, P., Tipping, M., Abed, L. and Veraksa, A. (2008) Tandem affinity purification in Drosophila: The advantages of the GS-TAP system. Fly (Austin) 2.

Lin, C.Y., Chen, S.H., Cho, C.S., Chen, C.L., Lin, F.K., Lin, C.H., Chen, P.Y., Lo, C.Z. and Hsiung, C.A. (2006) Fly-DPI: database of protein interactomes for D. melanogaster in the approach of systems biology. BMC Bioinformatics 7 Suppl 5, S18.

Lin, M.F., Carlson, J.W., Crosby, M.A., Matthews, B.B., Yu, C., Park, S., Wan, K.H., Schroeder, A.J., Gramates, L.S., St Pierre, S.E., Roark, M., Wiley, K.L., Jr., Kulathinal, R.J., Zhang, P., Myrick, K.V., Antone, J.V., Celniker, S.E., Gelbart, W.M. and Kellis, M. (2007) Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res 17, 1823-36.

Loevenich, S.N., Brunner, E., King, N.L., Deutsch, E.W., Stein, S.E., Aebersold, R. and Hafen, E. (2009) The Drosophila melanogaster PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. BMC Bioinformatics 10, 59.

Luengo Hendriks, C.L., Keranen, S.V., Fowlkes, C.C., Simirenko, L., Weber, G.H., DePace, A.H., Henriquez, C., Kaszuba, D.W., Hamann, B., Eisen, M.B., Malik, J., Sudar, D., Biggin, M.D. and Knowles, D.W. (2006) Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol 7, R123.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P., Rana, D., Riley, T., Sullivan, J., Watkins, X., Woodbridge, M., Lilley, K., Russell, S., Ashburner, M., Mizuguchi, K. and Micklem, G. (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biol 8, R129.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R.,

Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D. and Venter, J.C. (2000) A whole-genome assembly of Drosophila. Science 287, 2196-204.

Pacifico, S., Liu, G., Guest, S., Parrish, J.R., Fotouhi, F. and Finley, R.L., Jr. (2006) A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for Drosophila. BMC Bioinformatics 7, 195.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A. and Cesareni, G. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6 Suppl 4, S21.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 63, 490-500.

Rhee, S.Y., Wood, V., Dolinski, K. and Draghici, S. (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9, 509-15.

Saalfeld, S., Cardona, A., Hartenstein, V. and Tomancak, P. (2009) CATMAID: Collaborative Annotation Toolkit for Massive Amounts of Image Data. Bioinformatics.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., Wan, K., Rubin, G.M. and Celniker, S.E. (2002a) A Drosophila full-length cDNA resource. Genome Biol 3, RESEARCH0080.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., Yu, C., Carlson, J., George, R., Celniker, S. and Rubin, G.M. (2002b) The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res 12, 1294-300.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E. and Rubin, G.M. (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3, RESEARCH0088.

Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E. and Rubin, G.M. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol 8, R145.

Veraksa, A., Bauer, A. and Artavanis-Tsakonas, S. (2005) Analyzing protein complexes in Drosophila with tandem affinity purification-mass spectrometry. Dev Dyn 232, 827-34.

Wuchty, S. (2002) Interaction and domain networks of yeast. Proteomics 2, 1715-23.

Ye, J., Chen, J., Li, Q. and Kumar, S. (2006) Classification of Drosophila embryonic developmental stage range based on gene expression pattern images. Comput Syst Bioinformatics Conf, 293-8.

Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. and Gerstein, M. (2004a) Genomic analysis of essentiality within protein networks. Trends Genet 20, 227-31.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004b) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14, 1107-18.

Yu, J., Pacifico, S., Liu, G. and Finley, R.L., Jr. (2008) DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics 9, 461.

Zhai, B., Villen, J., Beausoleil, S.A., Mintseris, J. and Gygi, S.P. (2008) Phosphoproteome analysis of Drosophila melanogaster embryos. J Proteome Res 7, 1675-82.

Figure legends

Fig. 1 -Proteotypic Peptides (PTPs) for targeted proteomics. The diagram exemplifies the type of experiments that can be performed to identify and quantify specific peptides from complex sample mixtures taking advantage of PTPs. PTPs corresponding to the proteins of interest that are to be quantified in the sample are searched in the PeptideAtlas and chemically synthesized. A known amount of these PTPs labeled with a specific tag is added to the mixture where the unknown peptides from the samples are labeled with a different tag (e.g., an isotopic tag). The mixture of peptides is subdivided into different fractions and analyzed by mass spectrometry. Given that the mass spectra of the PTPs searched and the exact amount added to the mixture are both known, the proteins of interest can be identified and properly quantified (by the calculation of isotope ratios).

Fig. 2 -Two different interfaces of the DroID web page. The interfaces show the confidence score of each interaction pair, the original source from which that interaction has been predicted and the interaction map.

Fig. 3 –FlyMine web page. The homepage of FlyMine is shown with some of the data sources and other information this database provides.

Figure 1

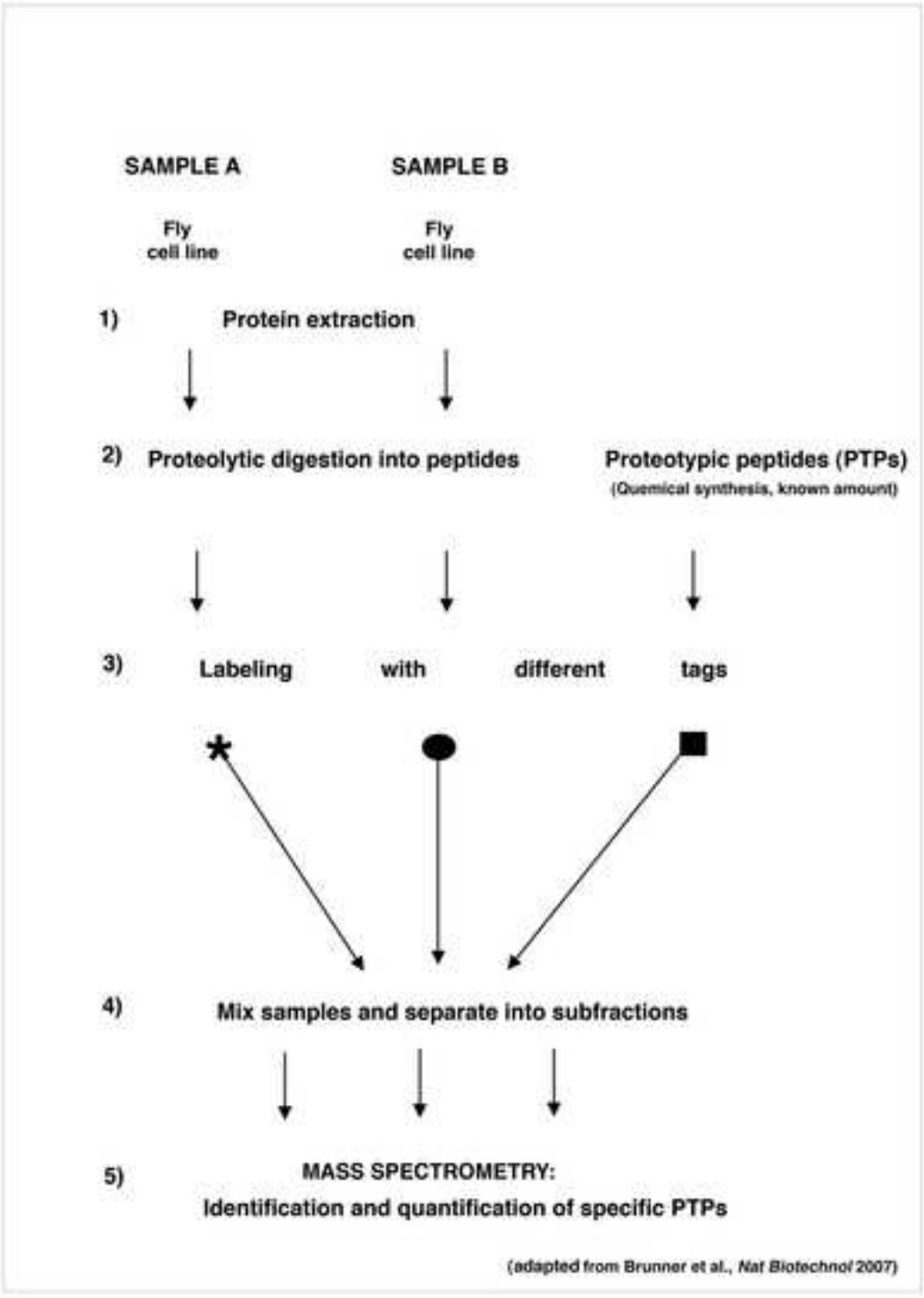

Figure 2

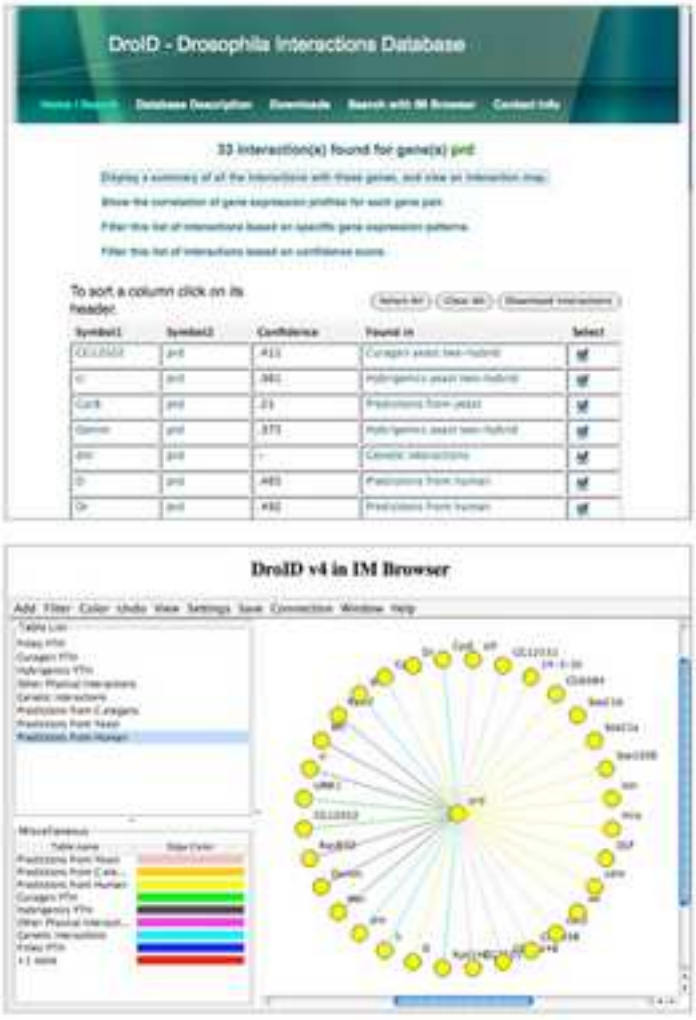

Figure 3

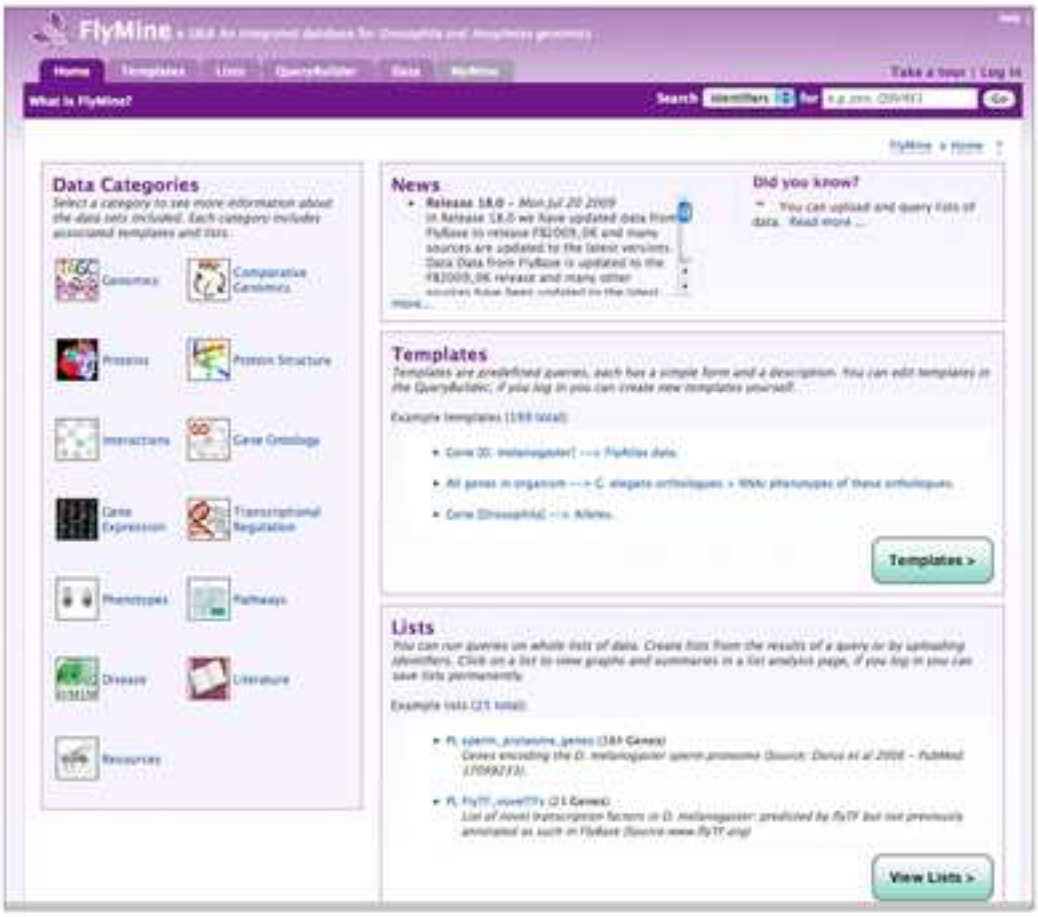

Table 1

Table 1: Summary of *Drosophila* proteomic tools and databases

PROTEOMICS ANALYSIS	RESOURCE	WEB PAGE
Protein identification, Quantification and PTMs	**PeptideAtlas**	http://www.drosophila-peptideatlas.org
	PhosphoPep	http://www.phosphopep.org
Protein-protein interactions (PPIs)	**Fly-DPI**	http://flydpi.nhri.org.tw
	Dro-ID	http://www.droidb.org.
	Hub protein classifier	http://www.cnbi2.ca/hub/
Protein expression patterns	***Drosophila* embryonic expression patterns**	http://www.fruitfly.org/cgi-bin/ex/insitu.pl
	CATMAID	http://fly.mpi-cbg.de/catmaid
Combined proteomic databases	**FlyMine**	http://www.flymine.org/
***Drosophila* primary database**	**FlyBase**	http://flybase.bio.indiana.edu/