BACKGROUND AND SIGNIFICANCE:

Summary: The study of model organisms has been central to reaching our current level of understanding of biological systems. Model Organism Databases (MODs) are the hubs of their respective research communities which number in the thousands worldwide. They provide a central location for genome sequence and annotation, gene expression, gene variation, interactions, phenotypic, functional and anatomical annotation as well as information on strains and reagents, the experimental literature and researchers. A current and future challenge for all the MODs is how to most usefully present the data they have compiled to the broadest user community in an integrated way: bench scientists have a growing need to access, query and manipulate large datasets and the more they can do this without the help of specialist bioinformaticians the better. There is a growing power in being able to carry out cross-organism comparisons, as the diversity and comprehensiveness of genomics dataset increases. Therefore there is a growing need to cross-compare and correlate data generated in different organisms. This proposal addresses the above needs for three of the leading model organism databases, representing *Saccharomyces cerevisiae* (SGD; Nash et al., 2007), *Rattus norvegicus* (RGD; Twigger et al 2007) and *Danio rerio* (ZFIN; Sprague et al., 2008). This will be done by extending the established InterMine data warehouse platform which provides a range of flexible and powerful features in a low-maintenance environment.

Manual curation of data is an essential function of the MODs and is of enormous value to the community. Over time the various MODs have evolved broadly similar working practices through responding to the needs of their respective users. However, the underlying infrastructure tends to have evolved separately so that the different resources present data in a way that is highly tuned to their community. Until now, the singular focus of MODs has tended to be towards presenting detailed information on single biological entities e.g. genes in the form of a report page with a rich array of annotation and links to other resources. Individual MODs have developed their own advanced query interfaces that are well-used but tend to be *ad hoc* systems without general flexibility and lacking the full power necessary to meet all the community's needs. That is, there is continuing demand from the research community for tools with which to fully query the MODs in complex ways and to be able to access bulk datasets derived from these queries in a flexible fashion. While the MODs have focused on curation of data and their presentation in a timely fashion to the research community, since its inception the focus of InterMine, as part of the FlyMine project (Lyne et al 2007, http://www.flymine.org), has been on the efficient integration, querying, list manipulation and display of diverse data types.

The recent and growing generation of high-throughput data has increased the challenges that the MODs face in providing such a comprehensive and integrated view of all available data to their respective research communities. Modern high-throughput data can be noisy, and to make it most useful it is important that the data are tightly integrated and can be queried alongside the high quality curated data assembled by the MODs. Incorporating new types of data, especially in large quantities requires continual development of the MOD database infrastructure leading to an additional call on limited resources. This is a task that InterMine has done for the FlyMine project and is now doing for a number of other projects, including the modENCODE project to identify all of the sequence-based functional elements in the *C. elegans* and *D. melanogaster* genomes (http://www.modencode.org). Indeed, InterMine has been designed to minimise the effort needed to incorporate new data types and also to deal effectively with large data volumes.

As yet there is little provision for researchers to make cross-MOD queries. The demand for this type of functionality is growing but currently such analysis typically requires the manual extraction of data and writing of custom analysis scripts. One of the features of this proposal is that we will address this issue of interoperability between the target MODs and beyond. By adopting a common platform we

can increase the ease with which data are exchanged between the different MODs to provide unique functionality for cross-organism comparison of data.

Examples of the kind of operation that have been demanded by the MOD user communities include: being able to extract the protein/ protein interaction complexes corresponding to genes with similar gene expression profiles; being able to query protein data and return lists of orthologous genes that are involved in human disease; having available additional data types that have not yet been modelled and incorporated within the MOD database environment. In addition, for ZFIN, the ability to incorporate and provide a query builder and batch download functionality and to include: protein (structure, function and pathway) data; homology data; disease data; and the Gene Ontology (The Gene Ontology Consortium, 2000) structured controlled vocabulary. For SGD, to include experimental results that provide a score for every nucleotide in the genome and not associated with a specific sequence feature; include PolII occupancy across the entire genome (e.g. Steinmetz et al., 2006) and transcript mapping (e.g. Miura et al., 2006) or ultra high throughput sequencing results. For RGD, users are frequently accessing information related to gene lists derived from such things as microarray studies, proteomic analyses, genes within rat QTL regions and orthologous genes within syntenic regions in mouse and human. We would like to make these list-based analyses much easier by allowing users to store their lists at RGD, and annotate their lists en masse using the wide variety of annotation within the database. Ultimately we would like to integrate more raw phenotypic data being generated by projects such as the Medical College of Wisconsin PhysGen project (http://pga.mcw.edu) which has genotyped 54 consomic (chromosome substitution) rat strains for over 200 phenotypes, accumulating a vast resource of data that need to be integrated with RGD and the genome. This proposal provides a way to address these demands. The MODs in this project receive hundreds of thousands to millions of quality visits per year, and millions of hits, indicating a strong demand for their services and the great potential impact of this proposal.

With increasing frequency bench biologists are confronted with the need to collect information or carry out analysis for hundreds or even thousands of genes, for instance when examining gene expression microarray data. Often this can involve the use of multiple web resources and it is common to find that there are naming inconsistencies between the platform used to generate the list of genes and the resource that is being used. This slows down work, as does the need to collate information from multiple sites into one table. Even if all the information needed is available at one site, it can still require a large effort to reformat as needed. InterMine can save a lot of time in these areas: it has a resilient list upload tool that can consider and provide feedback on resolving naming conflicts involving out-dated or ambiguous identifiers, and once uploaded the integrated analysis tools and widgets can be used without worry about naming conflicts. Furthermore, it is simple to configure custom tabular output (e.g. GO annotation, pathway annotation, phenotypes, disease association) and extract information on hundreds or thousands of e.g. genes in one step, something that otherwise can be extremely laborious.

In summary, there are a number of exciting advantages to adopting a common complex query platform: **1)** elements of the MOD user interface will become common across multiple organisms, reducing the barrier to use by a broader community of users; **2)** a common platform provides opportunities for interoperation, e.g. for comparative studies, that would be harder, if not too expensive, to achieve otherwise. Finally, **3)** by coordinating the effort in adopting a single generic system, InterMine, the MODs (and potentially other databases) can provide a complex query interface and additional useful functionality with much less effort than developing solutions independently. In addition, InterMine is part of the Generic Model Organism Database toolset (http://www.gmod.org); thus all the benefits of this development will be made freely available.

Although the period of support requested is short, the benefits of this proposal are long-term. The adoption of a common platform by three MODs at the same time makes good use of limited funds:

this proposal will provide the activation energy required to set up a new system and integrate it fully into the standard operating practices of each MOD. Furthermore, there is a growing community of InterMine database users and they will also benefit from the freely available developments made during this work. The fact that three model organism databases will adopt a common platform through this work will increase the potential for other model organism databases to follow suit. Therefore the impact of this project will extend long beyond the duration of the funding.