Phosphoregulators: Protein Kinases and Protein Phosphatases of Mouse

Alistair R.R. Forrest,1,2,3,9 Timothy Ravasi,1,2,4 Darrin Taylor,1,2,3 Thomas Huber,2,5 David A. Hume,1,2,3,4 RIKEN GER Group6 and GSL Members,7,8 and Sean Grimmond1,2

1The Institute for Molecular Bioscience, 2University of Queensland, Queensland, Australia; 3The Australian Research Council Special Research Centre for Functional and Applied Genomics, University of Queensland, Queensland, Australia; 4Cooperative Research Centre for Chronic Inflammatory Disease, 5Computational Biology and Bioinformatics Environment ComBiNE, 6Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-Ku, Yokohama, Kanagawa, 230-0045, Japan; 7Genome Science Laboratory, RIKEN, Hirosawa, Wako, Saitama 351-0198, Japan

With the completion of the human and mouse genome sequences, the task now turns to identifying their encoded transcripts and assigning gene function. In this study, we have undertaken a computational approach to identify and classify all of the protein kinases and phosphatases present in the mouse gene complement. A nonredundant set of these sequences was produced by mining Ensembl gene predictions and publicly available cDNA sequences with a panel of InterPro domains. This approach identified 561 candidate protein kinases and 162 candidate protein phosphatases. This cohort was then analyzed using TribeMCL protein sequence similarity clustering followed by CLUSTALV alignment and hierarchical tree generation. This approach allowed us to (1) distinguish between true members of the protein kinase and phosphatase families and enzymes of related biochemistry, (2) determine the structure of the families, and (3) suggest functions for previously uncharacterized members. The classifications obtained by this approach were in good agreement with previous schemes and allowed us to demonstrate domain associations with a number of clusters. Finally, we comment on the complementary nature of cDNA and genome-based gene detection and the impact of the FANTOM2 transcriptome project.

[Supplemental material is available online at www.genome.org.]

Regulation of protein activity by reversible phosphorylation, so-called phosphoregulation, is an important post-translational control mechanism implicated in many areas of biology. In this work, we use the term phosphoregulators to refer to both the protein kinases and the protein phosphatases. These enzymes regulate the phosphorylation status of the protein complement of a cell, and in turn, regulate the activity of their target phosphoproteins in cellular processes. The protein kinases covalently attach a phosphate group to a target, whereas the protein phosphatases remove these groups. Defining the entire complement of these proteins in the mouse gives us an opportunity to view the system as a whole.

The phosphorylation status of a protein, or more specifically, the pattern of phosphorylation on a given protein can determine its activity. The presence or absence of a phosphate group can change the conformation of the target protein, thereby modifying its activity. Phospho motifs in some cases provide binding sites for interactors, such as the 14-3-3’s (Yaffe2002). Alternatively, phosphorylation provides binding sites for enzymes catalyzing secondary modifications, such as further phosphorylation, dephosphorylation, acetylation (HIPK2; Hofmann et al. 2002), or in the case of ubiquitin ligases, facilitate ubiquitination and targeting of the protein for proteolysis (Ding and Dale 2002).

Phosphorylation events often occur in a cascade, in which activity of one kinase or phosphatase is dependent on the upstream activity of another. One of the best-studied examples of this is the regulation of the mitogen activated protein kinase (MAPK)-signaling cascade. MAPK signaling has no fewer than five levels of kinase regulation, MAP4K, MAP3K, MAP2K, MAPK, and MAPKAPK (Larochelle and Suter 1995; Cobb 1999; Dan et al. 2001) and one level of phosphatase regulation (MKP) (Theodosiou and Ashworth 2002). Furthermore, there is considerable cross talk between signaling cascades involving other phosphoregulators (Lehman and Gomez-Cambronero 2002), resulting in a network of phosphoregulators rather than a linear cascade.

Phosphoregulation is implicated in many areas of biology; these include transcriptional control (HIPK2; Hofmann et al. 2002; Pierantoni et al. 2002), signal transduction (MAPK; Cobb 1999), regulation of the cell cycle (Cyclin dependent kinases, NIMA kinases, cdk25 phosphatases; Nigg 2001), immunoproliferation (CD45 phosphatase; Koretzky et al. 1991), development (wnt signaling, β-catenin; casein-
sequences such as multidomain proteins and partial sequences. TribeMCL has been reported to handle problem-se-
ning. TribeMCL has been reported to handle problem se-
BLASTPe-values), which is then subjected to Markovcluster-
BLASTP output. The BLASTP output is parsed by the Tribe part
upon probability and graph flow theory (Van Dongen 2000; Enright et al. 2002). Interestingly, there are some phos-
regulators that shuttle dynamically between the cytoplasm
and the nucleus. The shuttling of cdc25 phosphatases (Dav-
eseac2000) and cyclin/cdk complexes (Toyoshima et al. 1998) is dependent upon the phase of the cell cycle or induc-
tion of a checkpoint, and is dependent upon phosphorylation

Eukaryotic protein kinases fall into two major classes on
the basis of distinct substrate preferences as follows: (1) ser-
ine/threonine kinases, and (2) tyrosine kinases. Similarly eu-
karyotic protein phophatases are classified as (1) serine/
theophine, (2) tyrosine, or (3) dual-specificity phophatases. Dual-specificity phophatases are able to dephosphorylate ser-
ine, theophine, and tyrosine residues. Each of these classes has
well-conserved catalytic domains, and as such, these classes are
amenable to domain-based sequence mining. These pro-
teins are well represented within the InterPro database of pro-
tein-sequence domains (Apweiler et al. 2001; http://www.ebi.
ac.uk/interpro). InterPro contains 12 domain entries impli-
cated in protein-kine activity and 10 domain entries asso-
ciated with protein-phophatase activity (Table 1). In some
cases, these domains represent subclasses of more general do-
main predictions. For example, a protein kinase can contain a
eukaryotic protein kinase motif (IPR000719) as well as a tyro-
sine protein kinase motif (IPR001245). In other cases, the do-

catalytic phosphatase catalytic subunit.

Using the InterPro domains detailed in Table 1, we set out
to identify all potential protein kinases and protein phos-
phatases of mouse. To this end, we mined all publicly avai-
lable cDNA sequences (including the FANTOM2 set) and the
Ensembl gene predictions. Once defined, the protein phos-
phatase and protein kinase complements were subjected to
clustering, and the resulting clusters examined for functional
groups and domain associations. An alternative approach to
traditional multiple alignment was adopted for clustering of
these two classes.

A recent development has been the release of a new pro-
tein sequence-similarity clustering tool known as TribeMCL
(Enright et al. 2002; http://www.ebi.ac.uk/research/cgg/tribe).
TribeMCL uses Markov clustering (MCL), an algorithm based
upon probability and graph flow theory (Van Dongen 2000)
to assign proteins to a given class. Distances are calculated on
BLASTP output. The BLASTP output is parsed by the Tribe part
of the package to produce a matrix of protein similarities
(BLASTP e-values), which is then subjected to Markov cluster-
ing. TribeMCL has been reported to handle problem se-
quences such as multidomain proteins and partial sequences

Table 1. InterPro Domains Associated With Protein
Kinases and Phosphatases

<table>
<thead>
<tr>
<th>Domain ID</th>
<th>Domain Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPR000719</td>
<td>Eukaryotic protein kinase</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR002290</td>
<td>Serine/Threonine protein kinase</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR001245</td>
<td>Tyrosine protein kinase</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR00961</td>
<td>Protein kinase C-terminal domain</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR01426</td>
<td>Receptor tyrosine kinase class V</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR002373</td>
<td>cAMP-independent protein kinase</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR001824</td>
<td>Receptor tyrosine kinase class III</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR002011</td>
<td>Receptor tyrosine kinase class II</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR002374</td>
<td>cGMP-dependent protein kinase</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR003527</td>
<td>MAPK</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR000239</td>
<td>GPCR kin</td>
<td>Kinases</td>
</tr>
<tr>
<td>IPR002291</td>
<td>Phosphorylase kinase γ catalytic subunit</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000387</td>
<td>Tyrosine specific protein phosphatase and dual specificity protein phosphatase family</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000340</td>
<td>Dual specificity protein phosphatase</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000934</td>
<td>Serine/threonine specific protein phosphatase</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR01932</td>
<td>Protein phosphatase 2C domain</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000242</td>
<td>Tyrosine specific protein phosphatase</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR003595</td>
<td>Protein tyrosine phosphatase, catalytic domain</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000222</td>
<td>Protein phosphatase 2C subfamily</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR002115</td>
<td>Mammalian LMW phosphotyrosine protein phosphatase</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000106</td>
<td>Low molecular weight phosphotyrosine protein phosphatase</td>
<td>Phosphatases</td>
</tr>
<tr>
<td>IPR000751</td>
<td>MPI_Phosphatase</td>
<td>Phosphatases</td>
</tr>
</tbody>
</table>

(Enright et al. 2002). It is a very fast algorithm, and as such, is
amenable to large data sets. It has been applied to the draft
human and mouse genomes with great success (Lander et al.
2001; http://www.ensembl.org) and forms the basis of the
Ensembl gene family assignments.

Markov clustering (MCL) can be thought of starting with a
matrix in which every protein node is interconnected by a
probability of transition (moving from one node to another).
Nodes that are highly related have a high probability of trans-

dition; nodes that are dissimilar have a low probability of tran-
sition. Starting at any given node, a random walk to another
node has a higher probability of moving within the same
natural cluster than between clusters. During iterative rounds
of expansion and inflation, terms specific to MCL, the strong
connections (those with a high probability of transition) are
strengthened, whereas weak connections are weakened fur-
ther. Inflation effectively severs flow between clusters,
whereas expansion dissipates flow within clusters (Enright
2002). Increasing the inflation value can increase the severity of
pruning, and hence, lead to higher granularity of clusters.
Over several rounds of expansion and inflation, the matrix
reaches a steady state, in which further expansion and infla-
tion have no effect, and clusters have effectively been iso-

Previous attempts to classify protein kinases have relied on
multiple sequence alignment followed by hierarchical tree
generation (Kostich et al. 2002). The assumption behind such
an approach is that there is conserved sequence between all
members to be classified. In the case of the protein kinases,
this is the catalytic region. The boundaries of the conserved
catalytic region must be determined and an optimal align-
ment produced. Decisions must be made on how much of the alignment to use, how partial sequences will be handled, and whether to include sequences that do not align well. Sequences that do not share conserved sequence cannot be classified in this way. For example, the protein phosphatases represent multiple classes with distinct evolutionary origins. The tyrosine phosphatases use a different catalytic mechanism to the serine threonine phosphatases (Kerk et al. 2002). TribeMCL does not require or assume conserved sequence, and as such, is amenable to this multiple class problem.

In this work, we report on the use of TribeMCL clustering to classify the murine complement of phosphorylators. We also report on domains associated with each of the clusters. Finally, we discuss the impact of the FANTOM2 transcriptome data (FANTOM Consortium and the RIKEN GSC Genome Exploration Group 2002) on these two important classes of proteins and the use of TribeMCL as a useful tool for dissecting out distinct classes of proteins.

RESULTS AND DISCUSSION
Identification of Protein Kinases and Protein Phosphatases in Mouse

We used the Sequence Retrieval System (Etzold and Argos 1993; http://srs.ebi.ac.uk) to query two public protein databases, SWALL and IPI (International Protein Index). By use of this tool, mouse sequences annotated as containing the kinase and phosphatase InterPro domains detailed in Table 1 were extracted. The RIKEN FANTOM2 group produced a nonredundant protein set (RPS), which combines public sequences with high-quality representative sequences from the FANTOM2 data set. We again used InterPro domain annotations to extract sequences from RPS3 (nonredundant representative protein set 3, available from RIKEN, http://fantom2.gsc.riken.go.jp). Additional partial RIKEN sequences earmarked as novel kinases or phosphatases during the MATRICS curation phase, which failed to make it through to RPS3, were also included.

Structure-based superfamily predictions (http://scop.mrc-lmb.cam.ac.uk/scop; Structural classification of proteins home page) are also available for the FANTOM2 data, however, the SCOP predictions for the kinases and phosphatases are broader and less well defined than the InterPro predictions. The SCOP PK-like (protein kinase-like) prediction encompasses many protein kinases, but it also includes many related sequences that are not protein kinases. In this respect, the SCOP predictions represent a superset of those identified by InterPro. For this reason, the SCOP-based predictions were not used for identifying the phosphoregulators.

Mapping of the Sequences to a Common Identifier

To identify transcripts from the same gene, the sequences were mapped to a common identifier by use of a combination of approaches (Fig. 1). The majority of sequence entries within IPI contain a reference to an Ensembl gene identifier. They also contain cross-references to the original peptide sequences within SWISS-PROT, TrEMBL, and REFSQ. Similarly, for the RPS and SWALL entries, there are references to the original peptide entries. Mappings to an MGD locus (Blake et al. 2002, http://www.informatics.jax.org/) are also provided for some RPS and IPI entries. In cases in which entries from the different data sources shared original sequence entries, they inherited the Ensembl gene identifier and MGD locus. Sequences that could not be directly assigned to an Ensembl gene had their respective cDNA sequences extracted. These cDNA sequences were then compared with the Ensembl cDNA sequences (available for download at http://wwwensembl.org/Mus_musculus) using BLASTN. These alignments were inspected manually. Those with significant hits were assigned the respective Ensembl gene identifier. The remaining sequences were compared with the Ensembl Gene Scan predictions (available in the same directory). Significant hits were assigned the Genscan locus. Finally, any remaining sequences were assigned the MGD locus (if available) or the EST accession number from which they came. The gold standard for mapping such a group of sequences would be to map them to a genomic location with a given orientation. We provide genomic positions for the majority of the sequences, however, eight of the kinase-related sequences could not be mapped.

By use of our mock locus approach with a preference for Ensembl gene identifiers, it was possible to consider all sequences. This made it simple to compare genomic predictions with transcript data. In Supplementary Tables 1 and 2, we provide extensive cross-referencing for all genes identified. Where possible, we provide representative IPI, RPS3 (and RPS6.3), SWALL, RIKEN, Ensembl, and MGD identifiers. We also provide a representative accession number and genomic position.

Nonredundant Sets of Protein Kinases and Protein Phosphatases

By use of the mock locus assigned in the previous step, the data sets were cross-referenced. For cluster analysis, the combined data set was sorted by mock locus, and the longest representative sequence taken. Intermeshing the predictions from the various protein sets resulted in an estimated 561 candidate protein kinases and 162 candidate protein phosphatases. It is worth noting that only 11 kinase-related sequences and 1 phosphatase sequence failed to map to an Ensembl locus or Genscan prediction.

A total of 341 of the candidate protein kinase sequences mapped to an Ensembl gene, 77 of these are predictions by Ensembl with no supporting transcript evidence in mouse. Of the remaining sequences, nine mapped to Genscan predictions, six were assigned to MGD loci, and the remaining five were labeled with their respective accession numbers.

Similarly, 158 of the candidate protein phosphatase sequences mapped to an Ensembl gene, 19 of these are predictions by Ensembl with no supporting evidence in mouse. A further three mapped to Genscan predictions, and the remaining one was assigned to an MGD locus.

Considering only those sequences with transcript evidence, 104 of the 484 candidate kinases were only supported by RIKEN evidence. This represents a significant proportion of kinases confirmed by transcript (21.3%). Similarly, for the candidate protein phosphatases, 27 of the 143 were only represented by RIKEN (18.9%). These sequences are novel transcripts in the respect that they are the only publicly available transcript evidence for these genes. However, the term novel may be inappropriate, as some of these sequences have been public since the first phase of the FANTOM project (The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium 2001). To clarify the terminology, we will use the term novel transcripts to refer to genes
Figure 1  Pipeline for identifying protein kinase and protein phosphatase containing sequences.
in which there is no obvious homolog and for which the FANTOM sequences are the only source of transcript evidence. The term homolog is used to identify sequences in which there is an easily identifiable homolog, and for which FANTOM sequences are the only source of transcript evidence. Additionally, there are uncharacterized sequences from other sources, but for the purpose of this analysis, these are considered known.

Clustering of the Nonredundant Kinase and Phosphatase Sets
TribeMCL-based clustering tended to separate the data sets into a few large clusters and many smaller ones. The number and size of clusters (i.e., the granularity of the clustering) could be adjusted by changing the inflation value used within TribeMCL or by using different expectation value (e-value) cut off values during the BLASTP step. The inflation value (a value between 1 and 5, default of 3) is used during the iterative expansion and inflation steps of TribeMCL, the higher the value the more granular the result. Adjustment of the e-value is not recommended by the authors of the TribeMCL package, however, adjusting the inflation value alone did not provide sufficient granularity. Decreasing the granularity led to smaller distinct clusters being split and merged with multiple large clusters rather than all members of a given smaller cluster going to a given larger cluster. We experimented with different combinations of e-value cutoff and inflation value (data not shown), and in the final data presented here, we use an expectation value of e-10 and an inflation value of 5. Using an e-value restriction of e-10 effectively sets the probability of transition to zero for any pair of nodes in which the e-value was higher than e-10 (i.e., those sequence comparisons with low BLAST scores). These settings tended to generate a large number of smaller clusters, which may indicate fragmentation of larger clusters. It was decided it was better for these to be assigned their own class rather than risk splitting them inappropriately between the larger clusters. These settings also separated smaller classes with distinct biology such as the guanylyl cyclases, TGF-β-type receptor kinases, and myotubulins (Tables 2 and 3).

Cluster Analysis of Protein Kinases
Two major kinase clusters were identified, a serine/threonine kinase cluster containing 289 members, with 63 previously uncharacterized transcripts, and a tyrosine kinase cluster containing 123 members, with 11 previously uncharacterized transcripts. The remaining 149 sequences, of which 37 were uncharacterized previously, were split into 59 smaller clusters, including 30 singletons (Table 2). A number of these smaller clusters represent kinases and kinase-like proteins with specialized biology and domain architecture; these include a TGF-β receptor kinase cluster containing 13 members, 2 of which were uncharacterized previously (cluster 2.0), a casein kinase-1 cluster with 12 members and 4 previously uncharacterized transcripts (cluster 3.0), and a guanylyl cyclase cluster with 8 members and 2 previously uncharacterized transcripts (cluster 5.0).

To further segment the large serine/threonine kinase and tyrosine kinase clusters, we used CLUSTALV (Higgins et al. 1992) to produce multiple alignments, and then used the alignments to create neighbor-joining trees with 1000 bootstraps. The larger branches from these trees were then used to subclass these two major kinase classes. (To view where subclass were derived from, please refer to the trees in Supplementary Figs. 1–4). The serine/threonine group split into five major subclusters, as did the tyrosine kinases. These subclusters could be split further into subfamilies that share a conserved set of domains and represent known kinase families (Table 2).

We took two of these subclusters as examples to demonstrate the clusters and their domain associations. We considered cluster 0.0, the first subcluster of the serine threonine group and cluster 1.3 of the tyrosine kinases (Table 2).

Within cluster 0.0, there are nine subfamilies, six of which appear to have domains that define them. The nine subfamilies are the ribosomal s6 kinases, the protein kinase C’s, a rho/myotony group, an unknown group, the G protein coupled receptor kinases, the serum glucocorticoid kinases, the RAC/ AKT kinases, the cyclic nucleotide dependent kinases (cNMP; cAMP, and cGMP dependent), and the aurora-related kinases. The PKC sequences share a PKC domain (IPR002219), a C2 domain (IPR000008), and a PKN domain (IPR000861). The Rho/myotony group is defined by a PKC domain, Pleckstrin-like domain (IPR001849), and citron-like domain (IPR001180), a subset of the unknown1k group contains PDX domains. The RAC group contains a pleckstrin-like domain, and finally, the cNMP-dependent group contains a cNMP-binding domain (IPR000595). For domain predictions on individual sequences, please refer to Supplementary Table 4, which shows the clusters and all of the InterPro domains found within each of the sequences.

Within the 1.3 cluster of the tyrosine kinases, there are four subfamilies, the ephrin receptors, the Janus kinases (JAK), the focal adhesion kinases (FAK), and a less-well defined group (EPH-like). The ephrin receptors are characterized by four InterPro motifs, these are ephrin (IPR001900), SAM—Sterile Alpha Male (IPR0001660), FIII—Fibronectin type III repeats (IPR003961, IPR003962), and RTK-V—receptor-type kinase-V motifs (IPR001426). The EPH-like group consists of six members, two of which share the ephrin and RTK-V motifs of the ephrins, but lack the FIII or SAM domains. The Janus kinase group is defined by an SH2 (IPR000980) motif and a Band4.1 motif (IPR000299), and finally, the focal adhesion group contains a Band4.1 motif and a focal-adhesion targeting region (IPR005189) (Table 2).

Cluster Analysis of Phosphatases
As with the kinases, the phosphatase sequences produced 2 major clusters and 40 smaller clusters that included 19 singletons. The largest was a Tyrosine phosphatase cluster containing 42 members, 2 of which were novel transcripts; the next largest was a Dual specificity phosphatase cluster containing 25 members with 7 novel transcripts. The remaining 64 phosphatases contained 19 novel transcripts. The smaller classes include nine myotubulins, nine serine/threonine phosphatases, nine protein phosphatase 2Cs (two novel transcripts), and five protein tyrosine phosphatase 4a/MRPL (prenylated phosphatases) (Table 3).

Domain Associations Within the Kinases and Phosphatases
All sequences in this study had their corresponding InterPro domain annotations extracted. Supplementary Tables 3 and 4 provide extensive domain information and cluster ID for every sequence in this study. Tables 2 and 3 summarize the most common domain associations for each of the clusters.
Table 2. Summary of Protein Kinase Domain Containing Clusters Identified by TribeMCL

<table>
<thead>
<tr>
<th>cluster ID</th>
<th>riken</th>
<th>total</th>
<th>branch</th>
<th>cluster or branch description</th>
<th>Hanks' classification</th>
<th>domain associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>10</td>
<td>66</td>
<td>SGK, Ribos6, RAC, PKC</td>
<td>serum glucocorticoid kinase, Ribosomal S6 kinase, Protein kinase-C</td>
<td>AGC-OTHER, AGC-VI, AGC-III, AGC-II</td>
<td>Pleckstrin like PKC, C2 domain, PKN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cNMP, Rho/myotin</td>
<td>cAMP/cGMP dependent PKs, Rho/myotin</td>
<td>AGC-I, AGC-VII</td>
<td>cNMP binding PKC, Pleckstrin, citron</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, aurora, GPCR</td>
<td>aurora-related, G protein coupled receptor kinase</td>
<td>AGC-OTHER, OPK-I, AGC-IV</td>
<td>regulator of G protein, GPCR</td>
</tr>
<tr>
<td>0.1</td>
<td>25</td>
<td>94</td>
<td>death assoc, CAMKI</td>
<td>death associated kinases, calcium/calmodulin-dependent kinase type II</td>
<td>CAMK-I, CAMK-I</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, MAPK</td>
<td>pim, mitogen activated protein kinase</td>
<td>OPK-OTHER, CAMK-OTHER</td>
<td>UBQ assoc, kinase assoc</td>
</tr>
<tr>
<td>0.2</td>
<td>13</td>
<td>72</td>
<td>PKD/ULK, unknown, MAPK</td>
<td>tousled, mitogen activated protein kinase</td>
<td>AGC-II/OPK-V, OPK-VII/CAMK-I, CMGC-II</td>
<td>mapk</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CDK, GSK, CKII</td>
<td>cyclin dependent kinases, glycogen synthase kinase, casein kinase II</td>
<td>CMGC-I, CMGC-OTHER/III/IV</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>5</td>
<td>apop MAP3K, unknown</td>
<td>polo like kinases, mitogen activated protein kinase</td>
<td>OPK-IV, OPK-OTHER</td>
<td>polo box</td>
</tr>
<tr>
<td>0.4</td>
<td>13</td>
<td>52</td>
<td>NEK, DUSP, MAP2K</td>
<td>NIMA expressed related kinases, dual specificity mitogen activated protein kinase kinases</td>
<td>OPK-V, OPK-II</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PAK, GCK</td>
<td>p21 activated kinases, Germlinal center kinases</td>
<td>OPK-III, PAK, citron</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, MAPK</td>
<td>mitogen activated protein 4K</td>
<td>OPK-III, PAK, citron</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, MAPK, SRC</td>
<td>ste20 related?</td>
<td>OPK-III, PAK, citron</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, SRC</td>
<td>ste20 related?</td>
<td>OPK-III, PAK, citron</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0</td>
<td>8</td>
<td>EGF/UFO/c-met</td>
<td>EGF/UFO/c-met receptor kinases</td>
<td>PTK-OTHER/XXI/XII, SH2, SH3</td>
<td>Ig, Fibrinogen III</td>
</tr>
<tr>
<td>1.1</td>
<td>7</td>
<td>50</td>
<td>RIPK, MAP3K/MLK</td>
<td>Receptor interacting PK</td>
<td>OPK-XI, PTK-XIII/XVIII</td>
<td>Ankyrin</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown, TIE</td>
<td>tunica interna endothelial cell kinase</td>
<td>PTK-III/IV, SH2, SH3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>UMK, RAF, GFR</td>
<td>lim motif containing kinases</td>
<td>PTK-VIII, PTK-IV</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>0</td>
<td>27</td>
<td>SRC, TEC/BTK</td>
<td>VEGFR, FGFR, PDGF, SRC related</td>
<td>PTK-IV, PTK-IV</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>leukocyte insulin recep</td>
<td>SRC proto-onco</td>
<td>PTK-IV, PTK-IV</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>eph like eph receptors</td>
<td>Brunton’s tyrosine kinase</td>
<td>PTK-VII, PTK-VII</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>3</td>
<td>28</td>
<td>leukocyte insulin recep</td>
<td>insulin receptor and related kinases</td>
<td>PTK-VIII, PTK-XI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>eph like eph receptors</td>
<td>ephrin-type A and B receptors</td>
<td>PTK-IV, PTK-IV</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
A number of InterPro domains were associated with the eukaryotic kinases. Excluding the kinase domains used to identify these sequences, the next most common InterPro domains were SH2 (IPR000980), Immunoglobulin/MHC (IPR001452), and Fibronectin, type-III repeats (IPR003961). These fall into two major categories, receptor-type domains and nonreceptor-type domains. Additional receptor-type domains observed and associated with specific groups were the various classes of immunoglobulin-like domains, receptor-type kinase domains, and the ephrin receptor-associated domains. The nonreceptor group appears to be enriched for interaction domains, most notably the SH2, SH3, PDZ (IPR001478), and pleckstrin-like domains (IPR001849), as well as citron (IPR001180) and ubiquitin-associated domains (IPR000449).

Similarly, within the phosphatases, two major groups could be identified. The most common domains associated with the phosphatases were the Fibronectin type-III repeats (IPR003961/IPR003962), Rhodanese-like domain (IPR001763), immunoglobulin-type domains (IPR003006/IPR003599/IPR003598), Band 4.1 (IPR000299), MAM (IPR000998), and GRAM domains (IPR004182). The FIII, Ig, and MAM domains were all associated with Receptor phosphatase domains. The Rhodanese-like domain is found in cdc25 phosphatases (m-phase inducing phosphatases) and a subset of the Dual specificity phosphatases, the MAPK phosphatases. The Band4.1 motif is associated with cytoskeletal interactors and the GRAM domain is known to occur in myotubularins (InterPro at EBI; http://www.ebi.ac.uk/InterPro/).

Misclassifications and Related Gene Families Detected by TribeMCL

Perhaps the greatest issue when using domain-based prediction is the identification of false positives. We encountered a number of nonphosphatase and nonkinase proteins that contained kinase or phosphatase InterPro domains. Within TribeMCL, these sequences clustered separately as small distinct clusters.

Proteins that catalyze very similar reactions to protein phosphatases, such as the PIP3 phosphatases, the myotubularins (cluster 2), and the PTEN-related genes (cluster 6) (Taylor et al. 2000, Maehama et al. 2001), represent one class of misclassification. Similarly, more distantly related enzymes, such as the mRNA-capping enzymes, mRNA 5'-triphosphatases (cluster 9) (Changela et al. 2001), acid sphingomyelinase-like phosphodiesterases (cluster 8) (Testi 1996), RNA lariat debranching enzyme (Kim et al. 2000), CD73-5'nucleotidase (Airas et al. 1997), and MRE11A (Hopfner et al. 2001) were all identified as containing protein phosphatase domains. Withintheprotein kinases, the guanylyl cyclases were also identified (cluster 5) (Lucas et al. 2000).

With the recognition that the data sets contained protein kinase-like and protein phosphatase-like sequences, we divided the data sets into two classes of trust. Clusters are labeled as protein kinase/phosphatase or protein kinase/phosphatase-like. Clusters containing member sequences with direct evidence of protein kinase or protein phosphatase activity within the literature and no conflicting reports (as with the myotubularins) were labeled as containing protein phosphatase domains. Within the protein kinases, the guanylyl cyclases were also identified (cluster 5) (Lucas et al. 2000).

Table 2. Continued

<table>
<thead>
<tr>
<th>cluster ID</th>
<th>riken</th>
<th>total</th>
<th>branch or branch description</th>
<th>Hanks' classification</th>
<th>domain associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4</td>
<td>1</td>
<td>10</td>
<td>JAK, Janus kinase</td>
<td>PTK-VII/VI</td>
<td>SH2, Band4.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FAK, focal adhesion kinase</td>
<td>PTK-IX</td>
<td>Band4.1, focaladhesion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>unknown12k</td>
<td></td>
<td>Ig, Frizzled, RTK-II</td>
</tr>
<tr>
<td>Other small classes</td>
<td>2.0</td>
<td>1</td>
<td>TGF-B recep</td>
<td>TGF-B, activin, BMP</td>
<td>OPK-IX</td>
</tr>
<tr>
<td>3.0</td>
<td>1</td>
<td>12</td>
<td>CK1, casein kinase 1</td>
<td></td>
<td>TGF-B R, TGF-b GS</td>
</tr>
<tr>
<td>4.0</td>
<td>1</td>
<td>12</td>
<td>CLK, cdc-like kinase</td>
<td></td>
<td>OPK-IX</td>
</tr>
<tr>
<td>5.0</td>
<td>2</td>
<td>8</td>
<td>GC, guanyl cyclase</td>
<td></td>
<td>CMG-C-V</td>
</tr>
<tr>
<td>6.0</td>
<td>1</td>
<td>7</td>
<td>cdk4/6, cyclin dep kin 4 and 6</td>
<td></td>
<td>CMGC-I</td>
</tr>
<tr>
<td>7.0</td>
<td>1</td>
<td>5</td>
<td>cdc2-like</td>
<td></td>
<td>CMGC-I</td>
</tr>
<tr>
<td>8.0</td>
<td>3</td>
<td>5</td>
<td>IRAK, IL-1 receptor associated</td>
<td>OPK-X</td>
<td>CMG-C-V</td>
</tr>
<tr>
<td>9.0</td>
<td>1</td>
<td>4</td>
<td>WNK, PK, lysine deficient</td>
<td></td>
<td>OPK-IV</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4</td>
<td>cdk-like, cdk-like kinases</td>
<td></td>
<td>CMG-C-I</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>3</td>
<td>GAK, cyclin G associated kinase</td>
<td></td>
<td>GC, extracellular ligand binding recep</td>
</tr>
<tr>
<td>undefined</td>
<td>16</td>
<td>76</td>
<td>undefined</td>
<td>cluster &lt;3 members</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TOTAL 104 561

Cluster ID refers to clusters identified by TribeMCL and CLUSTALV branches. Hanks' kinase class is provided to demonstrate overall agreement between the clustering (Hanks and Quinn 1991). Clusters labeled as unknown represent clusters in which a common underlying biology could not be identified readily for a given cluster.
502 protein kinase, and 59 kinase-like sequences. When considering sequences for which transcripts have been detected, 13 of the 96 protein-phosphatase sequences and 87 of the 435 protein-kinase sequences are only supported by RIKEN transcripts; these represent 13.5% of all protein phosphatases and 20% of all protein kinases detected in the mouse transcriptome. Also, within the protein-kinase-like and protein-phosphatase-like groups, RIKEN transcripts were responsible for 34.7% and 29.8% of these sequences, respectively.

Evaluation of TribeMCL

Examining the clusters obtained by TribeMCL, we observed a number of smaller clusters that seem to have split from the larger clusters. The cyclin-dependent kinases CDK4 and CDK6 (cluster 6.0) cluster away from the remaining CDKs (subcluster 0.2); whether this represents the true situation is debatable. Similarly, a number of the smaller phosphatase clusters (clusters 10.0 and 11.0) seem to represent proteins with similar domains to larger clusters. There are two Protein phosphatase 2c-type clusters (4.0 and 11.0) and two serine/threonine-type clusters (3.0 and 10.0) (Table 3).

This suggests that the granularity settings used in our analysis may have been set too fine. However, this level of granularity has been important in separating misclassifications from genuine hits (see previous section). Sequences placed within the smaller clusters or classified as singletons represent small classes with distinct biology.

Comparison to Ensembl

As mentioned previously, we identified 561 protein kinase and 162 protein phosphatase, related sequences. A total of 541 of the protein kinase-related sequences and 158 of the phosphatase-related sequences mapped to Ensembl genes. A total of 77 of the Ensembl kinase-related genes are only represented by a genomic prediction. Conversely, 20 kinase-related sequences did not map to Ensembl genes. This indicates an advantage in using a transcriptome-based screen, it is not dependent upon gene predictions. As to whether the 77 Ensembl predictions represent true genes, we will have to wait for transcriptome data to confirm their expression. A list of the sequences that failed to map to an Ensembl gene is available as Supplementary Table 5.

Comparison to Previously Published Classification Schemes

The clustering used in this work, on the basis of whole protein homology, presents an alternative to traditional domain-based homology assignments. A recent global analysis of the human members of the protein-kinase superfamily identified

| Table 3. Summary of Protein Phosphatase Domain Containing Clusters Identified by TribeMCL |
|---------------------------------|--------|--------|---------------------------------|--------|
| Cluster ID | rik | all | branch | branch or cluster description | domain associations |
| Tyrosine phosphatase (40) | 0.0 | 0 | 15 | recep 1 | Receptor tyrosine phosphatases—group 1 |
| | 0.1 | 1 | 5 | recep 2 | Receptor tyrosine phosphatases—group 2 |
| | 0.2 | 1 | 7 | recep 3 | Receptor tyrosine phosphatases—group 3 |
| | 0.3 | 0 | 7 | non-recep 1 | Non-receptor tyrosine phosphatases—group 1 |
| | 0.4 | 0 | 8 | non-recep 2 | Non-receptor tyrosine phosphatases—group 2 |
| DUSP (25) | 1.0 | 2 | 7 | DUSP | DUSP |
| | 1.1 | 3 | 13 | MKP | MAPK phosphatase |
| | 1.2 | 0 | 5 | T-DSP | DUSP, Rhodanese like |
| Ser/Thr (9) | 2.0 | 0 | 9 | myotubulin | myotubulin PIP3 phosphatase |
| | 3.0 | 0 | 9 | ser/thr | Ser/thr phos |
| PP2c (9) | 4.0 | 2 | 9 | PP2c | Protein phosphatase 2c |
| | 5.0 | 0 | 5 | 4a/MPRL | phos 4a and prenylated |
| | 6.0 | 1 | 4 | PTEI like | PTPN3 phosphatase |
| | 7.0 | 3 | 4 | cdc14 | Acidic Sphingomyelinase-like phosphodiesterase |
| | 8.0 | 0 | 3 | sphingo | mRNA capping enzyme |
| | 9.0 | 0 | 3 | mRNA capping | mRNA capping enzyme |
| | 10.0 | 0 | 3 | ser/thr 2B | Ser/thr 2B phosphatase |
| Undefined | 11 | 2 | 3 | unknown pp2c | unknown |
| | 12 | 0 | 3 | cdc25/MP1 | cdc25 phosphatase/M-phase inducer phosphatase |
| | TOTAL | 27 | 162 |

Cluster ID refers to clusters identified by TribeMCL and CLUSTALV branches. Cluster 11 is labeled as unknown pp2C, members contain a pp2C domain but no common underlying biology could be identified for this cluster.
510 candidate kinase sequences (Kostich et al. 2002), we identified a similar number of sequences, 502 protein kinases, and 59 kinase-like sequences. The study by Kostich et al. (2002), as does most other analyses, uses phenogroups constructed using the catalytic domain of the kinase for the alignment and tree assignments. This approach captures relations dependent on the catalytic domain, however, it ignores the effect of other domains that may be important for the biology of the whole protein. Perhaps surprisingly, we obtained similar classifications to those obtained by Kostich et al. (2002) and the earlier Hanks' classification scheme (Hanks and Quinn 1991; Protein Kinase Resource http://pkr.sdsc.edu/html/pk_classification/pk_catalytic/pk_hanks_class.html).

The most notable difference was the splitting of the cyclin-dependent kinases. Both the Hanks’ scheme and that of Kostich et al. (2002) place the CDKs together, however, the TribeMCL clustering separated the major CDK group (placed within cluster 0.2 of the serine threonine kinases) from three smaller CDK-related clusters (clusters 6.0, 7.0, and 10.0). The reasons for this difference are not clear, however, the other small TribeMCL clusters (2, 3, 5, 8, 9, and 11) represent proteins with specialized roles (Table 2). The fragmentation of the CDKs may reflect an underlying difference in their biology, however, an artefact of the clustering cannot be ruled out. These clusters contain a large number of genomic predictions. In cluster 6.0, four of the seven genes are predicted. Similarly, in cluster 7.0, three of five genes are predicted. In both cases, these are predicted genes from the Ensembl family ENSMUSP00000000708. Whether these predictions are somehow skewing the clusters is unknown.

For the mammalian protein phosphatases, a similarly comprehensive classification scheme is not available. A recent whole-genome analysis on the protein phosphatase catalytic subunits of Arabidopsis identified 112 candidate phosphatase sequences (Kerk et al. 2002); these were split into 69 pp2c, 18 DUSP, 23 serine/threonine, 1 tyrosine, and 1 low molecular weight tyrosine phosphatase. The small number of tyrosine phosphatases in Arabidopsis does not reflect the situation in mouse in which the tyrosine phosphatases represent the largest class of phosphatases. To assess the quality of the clustering for the tyrosine phosphatase clusters, we compared them with those identified by Andersen et al. (2001) (http://science.novonordisk.com/PTP/database.asp). There was good overall agreement with clusters separating into receptor type and nonreceptor type phosphatases, however, both cluster 0.2 and 0.4 contained both receptor and nonreceptor-type members.

Mouse Phosphoregulators

An advantage of examining both the protein kinases and protein phosphatases of mouse was to glimpse a global view of how these enzymes, which catalyze opposing reactions, could operate throughout the cell and how they are used in so many control systems. The most obvious observation was the high number of kinases in comparison to phosphatases. This is also the case in other eukaryotes, including yeast and man. A number of workers have commented on this previously, with the suggestion that protein kinases have higher specificities than protein phosphatases, however, this has been challenged (Zhang et al. 2002). There is clear evidence of substrate specificity by some phosphatases, consider the MKPs (MAP kinase phosphatases), which are specific for MAP kinases, and the cdc25 phosphatases specific to cyclin/CDK complexes.

The simplest explanation for this observation, however, is that a phosphatase is only required in systems in which the target protein needs to return to an unphosphorylated state. As mentioned in the introduction, there are alternative control mechanisms such as secondary modifications, which can modulate the activity of a phosphorylated protein, or in the case of ubiquitination, target it for destruction. Examples in

Novel Phosphoregulators Within the FANTOM2 Libraries

To assess the impact of the RIKEN FANTOM2 sequences, we only considered those for which the FANTOM sequence was the only source of a predicted peptide. Many of the other sequences presented here are also identified by RIKEN, but there is supporting evidence from another source. Some of these are also novel sequences, however, for the purpose of simplifying this assessment, these are considered as separate.

Starting with the phosphatase domain-containing sequences, we identified 8 homologs and 19 novel transcripts (Table 3). Novel transcripts for one MAPK phosphatase and two HSSH/slingshot-related dual-specificity phosphatases were identified in cluster 1. A transcript for a novel pp2c type phosphatase was identified in cluster 4. We also identified three smaller phosphatase-like clusters containing novel transcripts.

Cluster 6 contains PTEN (phosphatase and tensin)-like proteins. PTEN is an important PI3 phosphatase. The cluster contained PTEN and an additional three proteins, one prediction for a cyclin G-associated kinase (human GAK has a tensin and phosphotyrosine motif—O14976), one known, annotated as tyrosine phosphatase and one novel transcript with tensin homology. As to whether these constitute members of the PTEN family is not clear.

Cluster 7 contains four cdc14 protein tyrosine phosphatase sequences, one of these is a predicted gene by Ensembl, one is a homolog of a human gene, and two are novel transcripts. Cluster 11 contains three sequences of unknown function, they all contain a protein phosphatase-2C motif, they cluster separately from the pp2c phosphatase cluster 4 and separately from the structurally similar pp2c-like, pyruvate dehydrogenase phosphatase cluster 17. The remaining novel transcripts include a protein that clusters with the ecto-5’-nucleotidase, CD73 (cluster 15), and two transcripts with similarity to dual-specificity phosphatase DUSP13.

Within the serine/threonine kinase cluster, there were 62 RIKEN-only sequences, 28 of these are completely novel transcripts. These novel transcripts include a doublecortin domain-containing kinase, SNF1-related kinases, calcium/calmodulin-dependent kinases, MAPK-activated protein kinases (cluster 0.1), nima-related kinases, and p21-activated kinases (cluster 0.4). Within the tyrosine kinase cluster, there were 11 RIKEN only sequences, 6 of which are novel transcripts, including an ephrin-like kinase (cluster 1.3).

Within the remaining sequences, there were a number of clusters containing homologs and novel transcripts. Clusters and novel transcripts of note include a casein-kinase 1 cluster (3), containing 12 CK1-related sequences; we identified three homologs not observed previously in mouse, a novel kinase related to Elf-2α kinase (cluster 20), and an Il-1 receptor-associated kinase (IRAK) cluster (8) containing a hypothetical predicted by Ensembl, two novels, a homolog, and a known.
which dephosphorylation is likely to be important are ones in which a protein has a long half life, and during that half life, it needs to cycle dynamically between phosphorylated and unphosphorylated forms; alternatively, cases in which there is a need for very fine grain control, opposing kinases, and phosphatases may compete to determine the activity of a given protein. An example of this is the competing activities of Wee1 kinase and cdc25B phosphatase on mitotic cyclin/cdk complexes (Russell and Nurse 1986).

Conclusion
The FANTOM2 project highlights the advantages of a transcriptome-based approach. FANTOM2 has been able to confirm transcript prediction from the genome, but additionally, it has identified transcripts missed by the Ensembl gene-prediction pipeline. Some cases confirmed a Genscan prediction, whereas others identified completely undetected transcripts. Gene prediction algorithms applied to genomic sequence generally depend on a gene model, transcripts that break the model enhance our understanding of genes and how to better model them. This highlights the complementary nature of the transcriptome and the genome. Transcriptional evidence is necessary for defining coding regions within the genome and confirming genomic predictions. Conversely, genomic sequence is necessary for identifying potential transcripts and determining gene structure, in particular, exon-intron boundaries and promoter sites.

By mining sequence annotations for specific InterPro motifs, we have identified 561 candidate protein kinases and 162 protein phosphatases. Using TribeMCL protein sequence similarity clustering, we were able to separate the different classes of protein kinase and phosphatase within the sets. TribeMCL also provided a certain level of quality control to the classes assigned by InterPro motif detection, by allowing us to distinguish true kinase and phosphatase members from related gene family members. This separated the sets into 502 likely protein kinases and 96 likely protein phosphatases. Finally, the FANTOM libraries have provided us with the only transcript evidence for 13.5% of the protein phosphatases and 20% of the protein kinases described in the transcriptome. This represents a great resource, and the availability of full-length cDNAs from the FANTOM libraries will provide valuable clones necessary for functional confirmation of these genes in the future.

METHODS
Nonredundant Protein Sequences Used in This Analysis: RPS3, SWALL, and IPI
The RTPS group at RIKEN produced a nonredundant Representative Protein Set RPS3 that incorporates sequences from the public domain and the RIKEN FANTOM2 libraries. During the course of this study, the RPS set has undergone a number of updates; it currently stands at RPS 6 (http://fantom2.gsc.riken.go.jp). SWALL and the International Protein Index (IPI; http://www.ebi.ac.uk/IPI/) represent nonredundant protein sets available from EBI (http://srs.ebi.ac.uk). IPI is currently available for mouse and human, and merges entries from Ensembl, SWISS-PROT, TrEMBL, and REFESEQ. There is a large overlap between IPI and SWALL, however, SWALL has more partial sequences and does not incorporate Ensembl gene predictions.

Identification of Protein Kinases and Protein Phosphatases Within the FANTOM2 EST Sequences
Preceding the nonredundant kinase and phosphatase sets we detail in this work, a primary analysis was carried out on the raw FANTOM2 data that indicated a number of novel kinase and phosphatases were present. All sequences predicted by InterPro as containing a protein-kinase domain or a protein-phosphatase domain (Table 1) were inspected manually within the MATRICS annotation viewer and checked for similarity to known proteins. A small number of partial sequences identified in the curation phase were excluded from RPS3 (RPS had a requirement that the sequences be full length). These sequences were added back into the analysis prior to mapping.

Mapping of Mouse Kinase and Phosphatase Sequences to a Mock Locus
As detailed in the Results section, sequences identified from the nonredundant protein databases, SWALL, IPI, and RPS3, were merged. Where possible, sequences were assigned Ensembl identifiers or MGI loci. These were extracted from cross-references found in the sequence entry, and if entries from different databases were derived from the same SWISS-PROT, TrEMBL, or REFESEQ entry, they also inherited the Ensembl or MGI locus. Those sequences for which there was no mapping information had their underlying EST sequence extracted. This EST sequence was then used in a BLASTN alignment with the Ensembl gene predictions. The best three hits for each EST were examined manually. In cases in which the alignment was convincing, the sequence entry was assigned the Ensembl gene to which it hit. Similarly, those sequences that failed to hit an Ensembl gene were used in a BLASTN alignment of the Ensembl Genscan predictions. Those sequences that failed to hit either, were assigned the original identifying EST accession number. The Ensembl sequences used are available for download from http://www.ensembl.org/Mus_musculus. The standalone version of BLAST used was BLASTALL available at http://www.ncbi.nih.gov/BLAST.

Selection of Representative Sequences for Downstream Analysis
After assigning all of the sequences to a mock locus, it was possible to sort the sequences and choose a representative. The longest peptide sequence that mapped to a given locus was extracted for cluster analysis. Annotations were examined from all sequences that mapped to the same locus. If the annotations generally agreed, the most informative annotation was taken.

Domain Architecture
InterPro domain predictions for all sequences used in this analysis were extracted from the database from which the sequence originated. Domain predictions on sequences from the SWALL and IPI sets were extracted using the sequence retrieval system SRS at EBI (Etzold and Argos 1993; http://srs.ebi.ac.uk).

TribeMCL Clustering
The representative sequences for each family were clustered using TribeMCL (Enright et al. 2002; http://www.ebi.ac.uk/research/egg/tribe). The sequences were compared against themselves using an all-against-all BLASTP comparison. In this work, we used an expectation value cutoff of e-10 and used the BLOSUM62 matrix. The results of this blast comparison were then parsed and clustered by TribeMCL. We used an inflation value of five within TribeMCL to increase the granularity of the classifications. Mapping of cluster results to de-
scriptions and domain combinations was carried out in Microsoft Excel.

**CLUSTALV Clustering of Large TribeMCL Clusters**

CLUSTALV (Higgins et al. 1992) was used to segment the serine/threonine kinase cluster, the tyrosine-kinase cluster, the tyrosine-phosphatase cluster, and the dual-specificity phosphatase cluster produced by TribeMCL. CLUSTALV was first used to produce multiple alignments for sequences within the larger clusters; it was then used to create phylogenetic trees using the Neighbor joining option, with 1000 bootstraps. The larger branches within these trees were then used to subclass the TribeMCL clusters.

**Database Versions**

Ensembl mouse release (v. 7.3b3 12 July 2002). SWALL is updated weekly and IPI is updated monthly, the versions used to produce the nonredundant protein-kinase and protein-phosphatase sets were indexed the week of June 28 2002. RPS3—representative protein set 3 (http://fantom2.gsc.riken.go.jp).

**InterPro Assignments**

As detailed within the results, InterPro assignments for the initial identification of candidate sequences and for the later domain associations were extracted from the previously annotated data sets. This includes IPI, SWALL, FANTOM2, and RPS3. These assignments had been made using InterProScan, which are detailed at http://www.ebi.ac.uk/interpro.

**ACKNOWLEDGMENTS**

We thank the RIKEN Genome Exploration Research Group Phase I & II Team, Genomic Sciences Center, RIKEN, and the FANTOM consortium members. The Representative Protein Set RPS3 used in these analyses was generated by the RTPS group (RIKEN and FANTOM). The InterPro predictions were performed by Alexander A. Kanapin of the European Bioinformatics Institute.

**REFERENCES**


WEB SITE REFERENCES

http://fantom2.gsc.riken.go.jp; Functional annotation of mouse, RIKEN.
http://pkr.sdsc.edu/html/pk_classification/pk_catalytic/pk_hanks_class.html; Hanks’ kinase classification at The Protein Kinase Resource.
http://science.novonordisk.com/PTP/database.asp; Structural and evolutionary relationships among protein tyrosine phosphatase domains.
http://scop.mrc-lmb.cam.ac.uk/scop; Structural classification of proteins home page.
http://srs.ebi.ac.uk; The Sequence retrieval system (SR56).
http://www.ebi.ac.uk/interpro; The InterPro home page.
http://www.ebi.ac.uk/ipf/; International protein Index.
http://www.ebi.ac.uk/research/cg/tribe; Protein sequence clustering—TribeMCL.
http://www.ensembl.org/Mus_musculus; Ensembl mouse genome home page.
http://www.informatics.jax.org/; Mouse Genome Informatics (MDG).
http://www.ncbi.nih.gov/BLAST; Basic local alignment search tool home page.

Received November 1, 2002; accepted in revised form February 19, 2003.