Exploration of the Cell-Cycle Genes Found Within the RIKEN FANTOM2 Data Set

Alistair R.R. Forrest, Darrin Taylor, RIKEN GER Group, et al.

Genome Res. 2003 13: 1366-1375
Access the most recent version at doi:10.1101/gr.1012403

Letter

Exploration of the Cell-Cycle Genes Found Within the RIKEN FANTOM2 Data Set

Alistair R.R. Forrest,[1,2,3,7] Darrin Taylor,[1,2,3] RIKEN GER Group[4] and GSL Members,[5,6] and Sean Grimmond[1,2]

[1]The Institute for Molecular Bioscience, University of Queensland, Queensland Q4072, Australia; [2]University of Queensland, Queensland Q4072, Australia; [3]The Australian Research Council Special Research Centre for Functional and Applied Genomics, University of Queensland, Queensland Q4072, Australia; [4]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan; [5]Genome Science Laboratory, RIKEN, Hirosawa, Wako, Saitama 351-0198, Japan

The cell cycle is one of the most fundamental processes within a cell. Phase-dependent expression and cell-cycle checkpoints require a high level of control. A large number of genes with varying functions and modes of action are responsible for this biology. In a targeted exploration of the FANTOM2–Variable Protein Set, a number of mouse homologs to known cell-cycle regulators as well as novel members of cell-cycle families were identified. Focusing on two prototype cell-cycle families, the cyclins and the NIMA-related kinases (NEKs), we believe we have identified all of the mouse members of these families, 24 cyclins and 10 NEKs, and mapped them to ENSEMBL transcripts. To attempt to globally identify all potential cell cycle-related genes within mouse, the MGI (Mouse Genome Database) assignments for the RIKEN Representative Set (RPS) and the results from two homology-based queries were merged. We identified 1415 genes with possible cell-cycle roles, and 1758 potential paralogs. We comment on the genes identified in this screen and evaluate the merits of each approach.

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

The cell cycle, the process by which a cell replicates itself, is a highly controlled process employing many regulatory mechanisms. The importance of strict cell-cycle control is apparent when disregulation occurs, as it does it in cancer. Medically, the study of cell-cycle genes is of interest for understanding the nature of a given tumor type, but perhaps more tangibly it offers potential targets for chemotherapy (Sampath and Plunkett 2001; Carnero 2002). The recently launched Cancer Molecular Analysis Project (CMAP) at the U.S. National Cancer Institute (NCI) lists cell cycle as the major CMAP ontology for their molecular targets (http://cmap.nci.nih.gov/Ontologies).

The recently completed mouse genome (Waterston et al. 2002) coupled with the RIKEN—Functional Annotation of Mouse FANTOM transcriptome project (Okazaki et al. 2002), has provided a huge amount of genetic information to the researcher. Assigning function to these genetic elements is the next great step. Similarity to genes of known function is being used to suggest roles for novel proteins. Similarity measures include protein sequence homology, domain-based predictions (Apweiler et al. 2001; http://www.ebi.ac.uk/interpro, The InterPro homepage), and structure-based predictions (Murzin et al. 1995; http://scop.mrc-lmb.cam.ac.uk/scop, Structural Classification of Proteins homepage).

These approaches suggest a function by similarity to another protein; however, they do not present the researcher with a global view of the biology, nor do they present the function within a structured syntax. The Gene Ontology proposed by the Gene Ontology Consortium (2001) attempts to assign proteins to a number of structured ontologies. The top-level hierarchy is split into three ontologies: biological process, molecular function, and cellular component. These ontologies are then divided into lower-level ontologies. Any given gene product can belong to one or many ontologies, dependent upon what is currently known of the protein.

In this paper we attempt to identify genes with a role in the cell cycle, corresponding to biological process gene ontology GO:0007049 (and sub-branches). Within the context of the three higher-level ontologies, cell cycle-related genes encompass many lower-level ontologies. Within the context of cellular component, cell-cycle proteins have many different subcellular distributions, which often change during the course of the cell cycle. Within the context of biological process, cell-cycle genes encompass DNA replication, conformation and integrity, cytoskeletal changes, organelle distribution and reassembly, and most importantly, cell-cycle controllers that orchestrate the entire process.

Within the context of molecular function, cell-cycle regulation and mechanics involve many different classes of proteins at different phases throughout the cell cycle. Regulation at the levels of transcription (Whitfield et al. 2002), translation (Groisman et al. 2002; Horton et al. 2002), phosphorylation (Nigg 2001), and targeted proteolysis (Peters 2002) are all used during the cell cycle.

In the first part of the present study, we used similarity searches directed towards whole proteins and conserved domains to identify novel members of two prototype cell-cycle families, the NIMA-related kinases and the cyclins. Focusing on one family at a time allowed us to evaluate each assignment on a case by case basis. This approach requires good knowledge of each family in question and the level of homology within the given family. Requiring an in-depth knowl-

[6]Takahiro Arakawa, Piero Carninci, Jun Kawai, and Yoshihide Hayashizaki.
[7]Corresponding author.
E-MAIL a.forrest@imb.uq.edu.au; FAX 61-7-3365-4388.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1012403.

13:1366–1375 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

edge of each family does not make this approach amenable to a global screen.

In the second part of this study we applied a BLASTP-based approach using a single threshold for all hits. Two sources of cell cycle-related bait sequences were collected to screen the RIKEN variable protein set (VPS). The first set of bait sequences (Bait1) was obtained by downloading sequences corresponding to all the genes assigned cell-cycle roles within the AMIGO gene ontology browser (http://www.godatabase.org/cgi-bin/go.cgi). The second set of bait sequences (Bait2) was obtained by using a keyword search for cell cycle-related terms within the high-quality nonredundant protein sequence databases SWALL and the International Protein Index (IPI; EBI; http://srs.ebi.ac.uk). The results from both BLASTP searches are presented and compared to the assignments made by the Mouse Genome Informatics (MGI) group for the RPS sequences (Okazaki et al. 2002).

RESULTS AND DISCUSSION

NIMA-Related Kinases (NEKs)

NIMA-related kinases (NEKs) are defined by the prototype mitotic regulator NIMA (never in mitosis A), a protein kinase identified in *Aspergillus nidulans*. This kinase was identified by its ability to complement the *nimA5* temperature-sensitive mutation that causes cells to block in late G2 (Osmani et al. 1987). A number of mammalian homologs have been identified; however, only three have been shown to have cell-cycle roles (*Nek2*, Ha Kim et al. 2002; *Nek9*, Roig et al. 2002; *Nek11*, Noguchi et al. 2002).

Using the sequence retrieval system SRS6 (Etzold and Argos 1993; http://srs.ebi.ac.uk), we used keyword searches to extract all publicly available NIMA-related kinase sequences, and then we used these as bait sequences to blast against the RIKEN VPS. All sequences with significant hits were examined manually and mapped to the mouse genome.

In total we identified 10 NIMA-expressed related kinases (Table 1). All of these were either known to or predicted by ENSEMBL (http://www.ensembl.org/Mus_musculus). mRNA sequences for eight of these had been observed previously in mouse (1, 2, 3, 4, 6, 7, 8, and 9/nercc1). The two others were the mouse homolog of *Nek11* and a gene predicted by ENSEMBL (ENSMUSG00000037738) but not previously observed as a transcript; for the rest of this paper, this transcript will be termed *Nek12*. *Nek12* is immediately downstream of *Nek3*, and seems to represent a gene duplication which is conserved in human (*Nek3*:ENSG00000136098, *Nek12*:ENSG00000165396). These appear to represent distinct genes, as no transcripts were identified which included exons from both *Nek3* and *Nek12*.

During the course of mapping the NEKs however, transcripts were identified which were able to bridge the gap between multiple ENSEMBL genes. In the case of *Nek9*, recently published as *Nercc1* (Roig et al. 2002), AJ489828

Table 1. NIMA-Expressed Related Kinases and Their Splice Variants in Mouse

Symbol[a]	VPS ID[b]	cDNA[c]	Length[d]	Domains present[e]	ENSEMBL MOUSE[f]	Chromosome[g]	Position
Nek1	PC3906.3	AK034754	302	kinase	ENSMUSG00000031644	8	60172055–60264003
	PC3906.2	AK031330	424	kinase + lysine rich			
	PA3906.0	S45828	774	kinase			
	PC3906.1	AK077047	941	kinase + lysine rich	ENSMUSG00000031644	8	60172055–60264003
					ENSMUSG00000037970	8	60269845–60291714
Nek2	PC3907.1	AK077627	366	kinase	ENSMUSG00000026622	1	193201229–193212725
	PA3907.0	U95610	443	kinase			
Nek3	PC6195.1	NP_035978	509	kinase	ENSMUSG00000031478	8	20858571–20894987
	PA6195.0	AF099066	511	kinase			
Nek4	PC6196.1	AK078809	744	kinase	ENSMUSG00000021918	14	26049834–26085981
	PB6196.0	NM_011849	792	kinase			
Nek6	PC7896.3	AK029266	305	kinase	ENSMUSG00000026749	2	38989294–39024394
	PB7896.0	NM_021606	313	kinase			
Nek7	PC7895.1	AK078639	250	kinase	ENSMUSG00000026393	1	13936116–139440087
	PB7895.0	NM_021605	302	kinase			
Nek8	PC2055.1	AK014546	291	kinase	ENSMUSG00000017405	11	78804211–78814753
	PB2055.0	NM_080849	698	kinase + Rcc1			
Nek9/ Nercc1	PC30616.0	BC024926	196	none	ENSMUSG00000021249	12	80001495–80005923
	PC30616.X	AJ489828	984	kinase + Rcc1	ENSMUSG00000021249	12	80001495–80005923
					ENSMUSG00000034290	12	80007619–80020593
					ENSMUSG00000034284	12	80025106–80036284
Nek11	PC22564.1	AK030186	454	kinase	ENSMUSG00000035032	9	106012522–106082223
	PC22564.0	AK030042	628	kinase			
Novel–Nek12	PC23442.1	AK054168	336	kinase	ENSMUSG00000037738	8	20838052–20850939
	PC23442.0	AK032672	614	kinase			

[a]Gene symbol from MGI.
[b]Variable Protein Set ID.
[c]cDNA accession numbers.
[d]Length of peptide in amino acids.
[e]InterPro domains.
[f]ENSEMBL mouse gene assignment.
[g]Genomic location.

bridges three ENSEMBL mouse genes (ENSMUSG21249, ENSMUSG00000034290, ENSMUSG00000034284). In another example, a longer form of *Nek1*, detected by RIKEN clone 4932438I04, bridges two ENSEMBL genes, ENSMUSG00000031644 and ENSMUSG00000037970.

Interestingly, a number of the NEK genes appear to encode long and short forms (Table 1). The short forms contain the kinase domain, and longer forms may contain extra domains, as is the case with the Rcc1 domain in *Nek8* and *Nek9*. Splice variants for both *Nek11* (Noguchi et al. 2002) and *Nek2* have been previously identified, and in the latter example were shown to exhibit different activities and expression patterns (Hames and Fry 2002).

Cyclins

Cyclins are the regulatory subunit of cyclin/cyclin-dependent kinase (CDK) complexes; these are the prime movers of the cell cycle. Specific subsets of these complexes are required during different phases, the best known example being cyclin B/CDC2 (CDK1), required during G2/M progression (Smits and Medema 2001).

In a strategy similar to that used for the NIMA-related kinases, we extracted bait cyclin sequences from the SWALL protein database using SRS6. In the case of the cyclins, we were able to query the database for sequences containing Interpro (Apweiler et al. 2001) domains associated with cyclins, cyclin (IPR004366), or cyclin c-term (IPR004367). It was also possible to directly query the FANTOM2 data for the presence of these domains.

During the course of the analysis, 38 potential cyclins were identified within the RIKEN VPS, using a combination of the domain-based querying and iterative blasts. Upon further analysis, 24 of these held up as known or likely cyclins (Table 2). Sequences that failed to carry on to further analysis include three members of the Rb family (p105, p107, p130), three members of the TFIIB family, a number of other unrelated cyclin box containing proteins, and CABLES, a CDK5 interactor.

The 24 cyclins include the known cyclins A1, A2, B1, B2, B3, C, D1, D2, D3, E1, E2, F, G1, G2, H, I, K, T1, T2, and ania-6a (L) and ania-6b. Also included is uracil-DNA glycosylase 2, noted as a cyclin by a number of workers (Murray and Marks 2001). Two potentially novel cyclins were identified, one most closely related to CYCJ, cyclin J of *Drosophila* (Finley et al. 1996), and the other related to the L-type cyclins ania-6a and -6b. Both of these have been observed in the EST evidence. The cyclin J homolog (AK052506) has been seen in human BA690P14.1/FLJ10895, and the novel cyclin L (AK007413) is represented by another EST in mouse (BC027022). A tree showing the relationship between these two novel cyclins and the known cyclins is shown in Figure 1.

Cyclin J was identified in *Drosophila* in a screen for Dm-Cdc2 (cdk) interactors (Finley et al. 1996). This cyclin has been shown to associate with cdk2 and play a possible role in the nuclear form of the cell cycle that occurs within a common cytoplasm (syncytium), in *Drosophila* early embryogenesis (Kolonin and Finley 2000). EST evidence suggests that the cyclin J homolog is mainly expressed in spleen.

Cyclin ania-6a, also referred to as cyclin L, was identified using a differential display screen in rat brain. Dopamine and glutamate have been shown to induce distinct splice forms of ania-6a, and these forms are able to associate with the orphan CDK, PITSLRE. A close relative was also identified in a database search, termed ania-6b (Berke et al. 2001).

All three L-type cyclins have distinct EST distributions. Although ania-6a and -6b are both reported as brain-specific, there a number of ESTs which indicate that they are expressed in tumors, especially mammary tumors. Ania-6a is also enriched in kidney, whereas ania-6b is enriched in retina. The novel L-type cyclin is represented by ESTs in thymus and bladder.

Another interesting observation when examining the cyclins was the identification of a number of pseudogenes. During the course of annotation and at the BLAST query stage, multiple copies of cyclins B1, B2, D3, L and the novel L were identified within the VPS. In some cases this may be an artefact of the clustering used; however, for two sequences it is clear that they represent processed pseudogenes of cyclins B1 and B2, in that they correspond to intronless versions of the B1 and B2 coding sequence (CDS) with frameshifts. Where it was not possible to discriminate between the two VPS sequences based upon their alignment against the mouse genome, the respective gene was assigned both VPS identifiers. The B1 pseudogene (2610510C05RIK) is transcribed in the sense orientation, maps to chromosome 14 (3696256–3698151), and has four frameshifts relative to the cyclin B1 CDS. The B2 pseudogene (AK048139) maps to chromosome 8 (71302891–71304001), is transcribed in the antisense orientation, and contains multiple frameshifts relative to cyclin B2 CDS. The degraded CDS is more similar to cyclin B2 from golden hamster (CYCLIN B2 [*Mesocricetus auratus*] – P37883), which perhaps suggests that this is an older pseudogene.

Early mapping studies identified 10 cyclin B-related sequences within the mouse genome (Lock et al. 1992) In the present study, we identified five which are transcribed, three containing functional CDS, and two transcribed processed pseudogenes (both of which have supporting EST evidence other than RIKEN). Other workers have identified pseudogenes of cyclins D2, D3, G1, and UNG (Xiong et al. 1992; Lund et al. 1996; Kimura et al. 1997), with a transcribed cyclin D2 pseudogene suggested as a marker for decreased ovarian function (Kimura et al. 1997). The presence of processed pseudogenes of the cyclins within the transcriptome raises questions about whether these noncoding RNAs play some role in cyclin regulation.

Large-Scale Identification of Candidate Cell Cycle-Related Genes

As outlined in the introduction and Methods section, two sets of bait sequences were produced to mine the VPS for cell cycle-related genes. The first set, "BAIT1" contained 1859 sequences that correspond to genes assigned a cell-cycle role by the Gene Ontology Consortium (2001). The second, "BAIT2" contained 4437 sequences which correspond to eukaryotic sequence entries within SWALL and the IPI which have a reference to cell cycle. These bait sequences were formatted as databases and then used in a BLAST query of the FANTOM2 VPS set.

From the BLAST queries, significant hits (e-30) were observed for 698 of the BAIT1 sequences and 1042 of the BAIT2 sequences. If we cross-compare the VPS entries for which a hit was observed, we find that 416 of the VPS hits are shared between the two queries (Fig. 2A). If we then overlay this with the predictions from the MGI (while only considering the higher-quality predictions, that is, those supported by evidence other than SCOP or InterPro), we find 158 VPS entries supported by all three predictions, 495 by two or more predictions, and 1415 predicted by at least one method. These sequences represent the best hits for each query sequence and are those most likely to be homologs of known cell-cycle genes.

If we extend our predictions to consider all sequences

Table 2. Cyclin Genes of Mouse

Symbol[a]	Description	VPS ID[b]	Protein accession[c]	Nucleotide accession	Length[d]	ENSEMBL gene[e]	Chromosome[f]	Position
CcnA1	cyclin A1	PC1172.1	4933406P14RIK	AK077114	421	ENSMUSG00000027793	3	55513691–55523271
CcnA2	cyclin A2	PC1173.1	B130012D23RIK	AK044924	422	ENSMUSG00000027715	3	36445101–36452177
CcnB1	cyclin B1	PB1174.0	NP_031655	NM_007629	430	ENSMUSG00000041431	13	98163773–98171907
	cyclin B1 pseudogene	PC60003.0	2610510C05RIK	2610510C05RIK	431	ENSMUSG00000041431	14	3696256–3698151
CcnB2	cyclin B2	PC1175.1	2810438P13RIK	AK076122	398	ENSMUSG00000032218	9	71022667–71036508
	cyclin B2 pseudogene	#N/A	#N/A	AK048139			8	71302891–71304001
CcnB3	cyclin B3	PC21546.0	CAC94916	AJ416459	238	ENSMUSG00000039664	X	3147531–3163688
CcnC	cyclin C	PC6896.3	2310034I23RIK	AK009615	304	ENSMUSG00000028252	4	21310260–21332992
CcnD1	cyclin D1	PC1176.1	5730552F24RIK	AK017834	295	ENSMUSG00000031071	7	135381270–135389916
CcnD2	cyclin D2	PC1177.1	1810059D21RIK	AK007904	156	ENSMUSG00000000184	6	128026669–128047570
CcnD3	cyclin D3	PA1178.0/ PC29928.0	P30282/ C920026J05RIK	M86183/AK083384	292	ENSMUSG00000034165	17	46621806–46625848
CcnE1	cyclin E1	PC1180.1	9430027E08RIK	9430027E08RIK	491	ENSMUSG00000002068	7	28627925–28636455
CcnE2	cyclin E2	PC1181.2	2510008E20RIK	AK010929	403	ENSMUSG00000028212	4	10675389–10688099
CcnF	cyclin F	PC1182.7	D430050H14RIK	AK085183	777	ENSMUSG00000035996	17	2341843–23442489
CcnG1	cyclin G1	PA1183.0	P51945	L49507	294	ENSMUSG00000020326	11	41079097–41085715
CcnG2	cyclin G2	PC1184.2	A630048E05RIK	AK041941	163	ENSMUSG00000029385	5	92942535–92951144
CcnH	cyclin H	PA8616.0	Q61458	AF287135	323	ENSMUSG00000021548	13	82301682–82325153
Ccnl	cyclin I	PC1185.2	A130017J05RIK	AK079476	377	ENSMUSG00000029382	5	92857953–92877393
	J-like*	PC32637.0	D430039C20RIK	AK052506	379	ENSMUSG00000025010	19	40544158–40553302
CcnK	cyclin K	PC1186.1	AAH27297	BC027297	206	ENSMUSG00000021258	12	102443711–102452394
CcnL	cyclin L/ania-6a	PC70541.1/ PB7529.0	D130043G23RIK/ NP_064321	AK051380/ NM_019937	532	ENSMUSG00000027829	3	66478330–66490378
Pcee-pending	cyclin ania-6B	PC7227.1	1810019L15RIK	AK007552	224	ENSMUSG00000029068	4	151263482–151275446
	L-like*	PC10625.0/ PC60129.0	AAH27022/ 1810009O10RIK	BC027022/AK007413	249	ENSMUSG00000037261	11	79391180–79391845
CcnT1	cyclin T1	PC1187.1	NP_033963	NM_009833	724	ENSMUSG00000011960	15	99455704–99480696
CcnT2	cyclin T2	PC13788.0	9930005L08RIK	AK036772	173	ENSMUSG00000026349	1	128569488–128598789
Ung	UNG 2	PC33252.1	D930033C21RIK	AK086507	352	ENSMUSG00000042417	13	110585514–110587664

[a]Gene symbol from MGI.
[b]Variable Protein Set ID.
[c]Protein accession and corresponding Nucleotide accession number.
[d]Length of peptide in amino acids.
[e]ENSEMBL mouse gene assignment.
[f]Genomic location.
*L-like and J-like cyclins identified in this study. Pseudogenes were determined by BLASTN against genomic sequence.

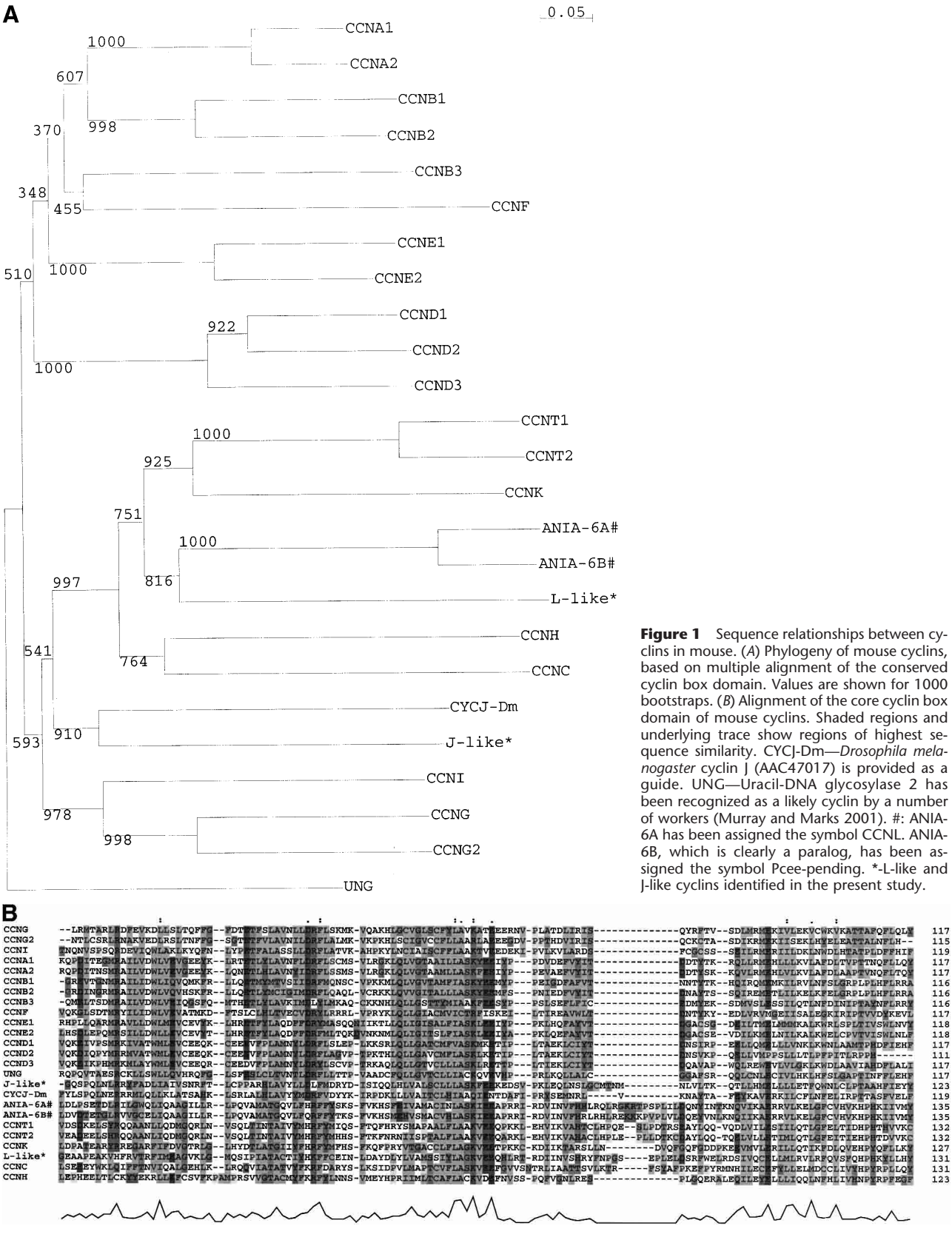

Figure 1 Sequence relationships between cyclins in mouse. (*A*) Phylogeny of mouse cyclins, based on multiple alignment of the conserved cyclin box domain. Values are shown for 1000 bootstraps. (*B*) Alignment of the core cyclin box domain of mouse cyclins. Shaded regions and underlying trace show regions of highest sequence similarity. CYCJ-Dm—*Drosophila melanogaster* cyclin J (AAC47017) is provided as a guide. UNG—Uracil-DNA glycosylase 2 has been recognized as a likely cyclin by a number of workers (Murray and Marks 2001). #: ANIA-6A has been assigned the symbol CCNL. ANIA-6B, which is clearly a paralog, has been assigned the symbol Pcee-pending. *-L-like and J-like cyclins identified in the present study.

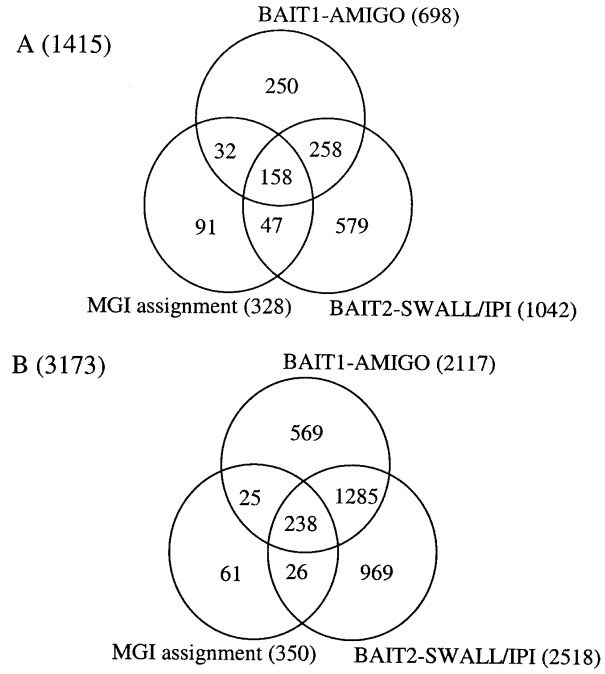

Figure 2 Distribution of candidate cell-cycle genes identified in the global screen. (*A*) Here, 1415 nonredundant sequences were identified as likely cell-cycle homologs by combining the best-hit predictions from the two BLAST searches and the MGI assignments. BAIT1–AMIGO corresponds to the (698) sequences identified by the bait sequences identified by the Gene Ontology Consortium as cell cycle-related. BAIT2-SWALL/IPI corresponds to the (1042) sequences identified by the bait sequences identified in a keyword search of SWALL and IPI. The 328 MGI assignments are those with good supporting evidence (InterPro and SCOP predictions were excluded). (*B*) Here, 3173 nonredundant sequences were identified when all significant hits were considered as well as the InterPro-based MGI assignments.

with a significant hit, such that a given bait sequence can have significant hits from more than one VPS sequence, we are able to identify potential paralogs of the genes identified in the first phase (Fig. 2B). When we consider all significant hits and also include the lower trust InterPro domain to GO mappings from MGI, we identify 3173 sequences which are identified by at least one source. These can be broken into the original 1415 candidate cell-cycle genes and 2468 potential paralogs.

Cell Cycle-Related Sequences Missed by This Approach

The e-value cutoff used in this study is likely to exclude a number of distantly related sequences. During the MATRICS curation phase of the FANTOM2 project, a number of cell-cycle families were investigated in detail. Within the raw FANTOM2 data, nine known histone deacetylases (HDACs) and two novels were identified (AK083724, AK045994); similarly, three known centrins and one novel (AK078275) were identified. These sequences served to evaluate first the RPS and VPS sets and secondly the large-scale BLAST approach used in this study.

All of the centrins identified in the FANTOM2 data were also present in the RPS set; however, two of the HDAC sequences were lost, one a known sequence, and one a novel sequence. The BLAST search identified the remaining known centrins and HDACs; however, none of the novel sequences for these two families were identified. This appears to be because the e-value cutoff was set too high; this would argue in favor of a lower threshold. This highlights the problem of requiring family-specific thresholds when evaluating multiple families. In the case of the novel HDAC (AK045994), an e-value cutoff of e-18 would have hit HDAC_MAIZE (probable histone deacetylase [RPD3 homolog]—*Zea mays*), and similarly the novel centrin (AK078275) would have been detected with a threshold of e-28, hitting CATR_SCHDU (caltractin [centrin]- *Scherffelia dubia*).

Reassuringly however, the e-30 threshold used allowed us to identify all of the cyclins and NEKs identified in the family-based screen. All members were identified in the paralog table; however, the best-hit table missed the novel *nek12*.

Due to the nature of keyword-based sequence extraction and the incomplete nature of the Gene Ontology Database, it is unlikely that we have captured all cell cycle-related genes. For a gene with known cell-cycle biology to be included in either set of bait sequences, either the biology must have been recorded in terms of gene ontology, or the sequence entry for that gene must contain cell cycle-related keywords. There are likely to be sequences that have evidence of a cell-cycle role reported in the literature but do not fit either of these criteria. Capturing such sequences requires a more complex text-mining-based strategy that would link sequences to evidence in the literature.

Sequences Falsely Assigned as Cell Cycle-Related

Two potential sources of false positives in this study arise from (1) the use of keyword searches to extract sequences containing cell cycle-related annotations, and (2) the e-value cutoff used. Keyword searches via SRS are not context-sensitive and consequently, there is a failure to discriminate between "… gene X is involved in cell cycle", "… gene X is possibly involved in cell cycle" and "… gene X is NOT involved in cell cycle". Bait sequences for which a significant hit was observed had their annotations checked manually for context. Cases where a keyword could be used in another context such as "purified 'metaphase' chromosomes"-[SWALL:'PYRG_HUMAN'] or "lipid-water 'interphase'" [SWALL:'GLUC_BOVIN'] were removed on a case by case basis. By careful manual curation, the impact of this was minimized.

The choice of e-value was set high to reduce false positives. As discussed in the introduction, the choice of an optimal e-value suitable for multiple families is difficult, as the level of homology within different families varies significantly. The homolog assignments took the best hit with a score below e-30. Varying the thresholds used for BAIT1 between e-30, e-50, e-80, and e-100 gave 698, 603, 501, and 450 hits, respectively. Similarly, varying the thresholds for BAIT2 resulted in 1042, 923, 786, and 700 hits, respectively. The e-values for every hit are provided in the Supplementary data so that researchers can apply a more stringent cutoff if desired.

The impact or e-value cutoff is more significant when considering paralogs. Large highly conserved families such as the zinc finger proteins and protein kinases are overrepresented in the paralogs table. As an example, 192 hits from BAIT2 are from ZF35_MOUSE (zinc finger protein 35), and the next most frequent is 37 hits to FER_HUMAN (proto-oncogene tyrosine-protein kinase FER). Similarly from BAIT1, 129 hits are from another zinc finger protein, Q9NZH2. If we limit the number of hits to the best hits up to a maximum of five, the numbers for the BAIT2 search drop from 2518 to 1851 and the BAIT1 search from 2117 to 1441.

Another source of overestimation is from bait sequences

annotated as "similar to cell cycle-related gene X". Similarly, sequences within GO with sequence homology-based evidence may be assigned a cell-cycle role. Thus we end up with a (similar to [similar to cell cycle-related gene X])-type association, where the bait has no direct evidence of a cell-cycle role.

Comparison to Functional Screens

In a recent microarray-based screen of the human cell line HeLa, 874 genes were identified as cell cycle-regulated (Whitfield et al. 2002). Similarly, in two independent microarray-based screens of budding yeast, 416 and 800 open reading frames (ORFs) of the 6220 monitored transcripts were identified as cell cycle-regulated (Cho et al. 1998; Spellman et al. 1998). As noted by those authors, their work identifies only genes that are regulated at the level of transcription.

The Munich information center for protein sequences (MIPS; Mewes et al. 2002) holds a database of ORF information for *S. cerevisiae*. A simplified gene ontology within MIPS, the 'functional catalog' (http://mips.gsf.de/proj/yeast/catalogues/) identifies 3936 ORFs with some assigned function, and 451 of these are identified as cell cycle-related. Also within MIPS, the yeast 'phenotype catalog' identifies 1469 ORFs associated with mutant phenotypes, 274 of which are cell cycle-related.

Comparing these estimates of the complement of cell-cycle genes within human and yeast, we see an estimated 3% of human genes and 6.7%–18.7% of yeast genes. The lower estimates are based solely upon microarray data and only identify those genes that are transcriptionally regulated with the cell cycle. The highest estimate of 18.7% is based on the observed yeast mutant phenotype. In the present study, we identified 1415 sequences as likely cell-cycle proteins; this represents 7.5% of the 18,768 protein coding genes identified in mouse (Okazaki et al. 2002).

Mouse Cell-Cycle Genes

Within the best hits there were 512 sequences where the best supporting evidence was from mouse. These included well known cell-cycle genes such as cyclin-dependent kinases (CDK), CDK inhibitors and interactors, cyclins MADs (mitotic arrest-deficient) and BUBs (budding uninhibited by benzimidazoles), cell division controllers (cdc), the E2F transcription factors, histone deacetylases (HDACs), mini-chromosome maintenance (MCMs), calmodulins, septins, RADs, cullins, kinesins, lamins, and a significant number of phosphoregulators such as the mitogen-activated protein kinases (MAPKs), NEKs, Polo-like kinases, CHKs and cdc25 phosphatases, as well as representatives from many other protein families.

Homologs with cell-cycle evidence from 53 other species made up the remaining 903 sequences. Most of these are also known to have cell-cycle roles in mouse; however, as the bait sequence from which the hit was made was not a mouse sequence, it must be inferred that the corresponding mouse entry was not detected by the Gene Ontology Consortium or by the keyword search. The majority of the remaining sequences came from *H. sapiens* (414), *S. pombe* (149), and *D. melanogaster* (70).

A summary table identifying the candidate cell-cycle sequences is provided as Supplementary Table 1, and includes the evidence for the prediction, and in the case of the BLAST predictions, the e-value and the species of origin, and in the case of the MGI assignments, the evidence and a trust assignment. Lastly, provisional Gene Ontology assignments are provided. Regarding the MGI predictions, these were directly from the MGI assignments. Regarding the sequences from the GO-based BLAST, the assignments were inherited from the bait sequence. The remaining sequences from the keyword search were provisionally assigned gene ontologies based on the keyword used to extract the bait sequence. The predicted paralogs are also presented as Supplementary Table 2.

Definition of a Cell-Cycle Gene

Evaluating any given gene for a role in the cell cycle based on homolog alone is useful but not completely conclusive. The cyclins demonstrate an important caveat when using homology-based evidence for assigning a role. The ania-6a cyclin clearly meets the definition of a cyclin in that it forms the regulatory subunit of a cyclin/CDK complex with PITSLRE. However, this cyclin has not been shown to have a cell-cycle role (ania-6a; Berke et al. 2001). Similarly, selected members of the 14-3-3 family of proteins have cell-cycle roles (Yaffe 2002), most notably RAD23 and 14-3-3 σ; however, not all members have been shown to have a cell-cycle role.

In most cases a family can be defined by a motif or shared homology. However, the presence of a domain does not necessarily infer a common biology; there are a number of cyclin box-containing proteins which do not have a cell-cycle role. Historically, families have also been assigned by the phenotype observed. This includes "families" such as the MADs and BUBs (Wassmann and Benezra 2001) and RADs (Rowley 1992) that on a domain basis represent multiple families.

The next point to consider is the definition of a cell cycle-related gene. Cell-cycle controllers are clearly cell cycle-related, and some of these are expressed in a phase-specific manner; however, the question remains as to whether expression in a phase-dependent manner constitutes a cell-cycle role. Differentiating between these biologies is possible using the structured ontologies presented by the Gene Ontology Consortium (http://www.geneontology.org/). The added benefit of a gene ontology assignment is the evidence code recorded. This gives us some feel of trust in the assignment.

To place a gene within the context of a cell-cycle role based upon homology alone is not enough; predictions need to be verified at the bench. In silico assignments suggest a role for a given gene, but on-bench confirmation is needed to put trust in the assignment. Combined, the in silico predictions guide us when deciding on which experiments to conduct at the bench, and the bench results give us feedback on the in silico model used.

Gene Ontology Associations

We next assessed the distribution of the various gene ontologies assigned by MGI to the RPS set and looked for associations with the cell-cycle genes identified in Figure 2A. Table 3 shows the top 30 assignments for each of the three top-level hierarchies (molecular function, biological process, and cellular component). For this analysis we ignored the more speculative predictions supported by SCOP.

As expected, cell cycle-related ontologies were enriched in the biological process branch; however, there were a few surprises. The most common ontology within this branch was protein amino acid phosphorylation, followed by regulation of transcription, then regulation of cell cycle. Similarly, within the context of molecular function, transcription factors and protein kinases featured heavily.

Within the context of cellular component, almost a third of the predicted peptides were assigned to the nucleus. The next most common assignments were intracellular location

Table 3. Top Thirty Gene Ontologies Associated With the Cell-Cycle Genes Identified in Figure 2A

Interpro ID	Molecular function	Count	Interpro ID	Biological process	Count	Interpro ID	Cellular omponent	Count
GO:0005524	ATP binding	401	GO:0006468	protein amino acid phosphorylation	170	GO:0005634	nucleus	422
GO:0003677	DNA binding	256	GO:0006355	regulation of transcription, DNA-dependent	164	GO:0005622	intracellular	133
GO:0016740	transferase	192	GO:0000074	regulation of cell cycle	129	GO:0016021	integral membrane protein	104
GO:0004672	protein kinase	169	GO:0007049	cell cycle	111	GO:0005615	extracellular space	80
GO:0004674	protein serine/ threonine kinase	169	GO:0008151	cell growth and/or maintenance	95	GO:0016020	membrane	75
GO:0004713	protein tyrosine kinase	158	GO:0016288	cytokinesis	66	GO:0005737	cytoplasm	40
GO:0003700	transcription factor	88	GO:0006281	DNA repair	64	GO:0005856	cytoskeleton	35
GO:0016787	hydrolase	84	GO:0006260	DNA replication	60	GO:0005875	microtubule associated protein	30
GO:0016301	kinase	83	GO:0007242	intracellular signaling cascade	59	GO:0005667	transcription factor complex	29
GO:0003676	nucleic acid binding	72	GO:0007165	signal transduction	45	GO:0005694	chromosome	21
GO:0005515	protein binding	51	GO:0006810	transport	41	GO:0005739	mitochondrion	18
GO:0005525	GTP binding	50	GO:0007017	microtubule-based process	41	GO:0005886	plasma membrane	18
GO:0005509	calcium ion binding	48	GO:0007264	small GTPase mediated signal transduction	36	GO:0005871	kinesin	17
GO:0003774	motor	41	GO:0007067	mitosis	35	GO:0005783	endoplasmic reticulum	16
GO:0004872	receptor	41	GO:0006508	proteolysis and peptidolysis	33	GO:0005840	ribosome	16
GO:0000166	nucleotide binding	40	GO:0006412	protein biosynthesis	31	GO:0005718	nucleosome	15
GO:0003723	RNA binding	39	GO:0006470	protein amino acid dephosphorylation	27	GO:0005882	intermediate filament	14
GO:0003685	DNA repair protein	33	GO:0016538	cyclin-dependent protein kinase, regulator	26	GO:0005887	integral plasma membrane protein	14
GO:0003924	GTPase	33	GO:0007275	development	24	GO:0005717	chromatin	12
GO:0004386	helicase	29	GO:0007001	chromosome organization and biogenesis (sensu Eukarya)	23	GO:0005794	Golgi apparatus	12
GO:0003925	small monomeric GTPase	28	GO:0008283	cell proliferation	23	GO:0016459	myosin	12
GO:0003779	actin binding	26	GO:0006886	intracellular protein transport	22	GO:0005576	extracellular	11
GO:0008026	ATP dependent helicase	26	GO:0015031	protein transport	22	GO:0005813	centrosome	9
GO:0005198	structural molecule	25	GO:0006464	protein modification	21	GO:0005891	voltage-gated calcium channel complex	9
GO:0003754	chaperone	23	GO:0008152	metabolism	20	GO:0008287	protein serine/threonine phosphatase complex	9
GO:0000158	protein phosphatase type 2A	20	GO:0006118	electron transport	19	GO:0005663	DNA replication factor C complex	8
GO:0000163	protein phosphatase type 1	20	GO:0006334	nucleosome assembly	18	GO:0005681	spliceosome complex	8
GO:0003777	microtubule motor	20	GO:0006511	ubiquitin-dependent protein catabolism	18	GO:0030286	dynein	8
GO:0003824	enzyme	20	GO:0006915	apoptosis	18	GO:0005578	extracellular matrix	7
GO:0004722	protein serine/ threonine phosphatase	20	GO:0007050	cell cycle arrest	18	GO:0005624	membrane fraction	7

Gene Ontology assignments from MGI for the RPS sequences were extracted and ranked. Assignments made by SCOP to GO mappings were ignored for this table.

and integral membrane proteins. Themes that ran through the data set include phosphoregulators (kinases and phosphatases), transcriptional regulation, cytoskeletal proteins, and membrane proteins.

Conclusion

Using a focused screen of the NIMA-related kinases and the cyclins, we identified all known members and a number of novels. We mapped these sequences to ENSEMBL gene predictions and made comments on their biology.

In the second part of the study we not only identified the majority of mouse genes with known roles in the cell cycle, but also significantly extended the number of cell-cycle assignments. We identified 1415 likely cell-cycle genes and a further 1758 paralogs. Within these two sets we identified novel members of known cell cycle-related families and homologs of cell-cycle genes from closely and more distantly related species.

The results of the global screen, the data presented in Supplementary Tables 1 and 2, represent the largest known assignment of cell-cycle genes within mouse and quite possibly within any organism. These entries are indexed by VPS and RPS peptide identifiers. The representative EST accession number, the tentative gene ontology assignments, and the evidence for each of the predictions are also provided. These assignments are provided to the research community as a resource for further investigation and experimental validation.

METHODS

RIKEN Data Sets Used

RPS and VPS are nonredundant sequences identified by the RIKEN RTPS group; these attempt to merge all publicly available, high-quality EST sequences with the RIKEN FANTOM2 sequences. The corresponding cDNA and protein sequences are referred to as the RTS (representative transcript set), RPS (representative protein set), and VPS (variable protein set). The VPS aims to produce a set of all unique alternative transcripts from a given transcriptional unit (Okazaki et al. 2002).

For the majority of the present study, the VPS was used. The VPS was used in preference to the RPS to capture cases where a VPS sequence had a better hit with a known cell-cycle regulator than the representative RPS sequence. The raw RIKEN FANTOM2 sequences were accessed using MATRICS (http://fantom2.gsc.riken.go.jp).

Publicly Available Databases Used

Mouse ENSEMBL cDNA sequences and GENSCAN predictions were downloaded from ENSEMBL (http://www.ensembl.org/Mus_musculus/). The sequence retrieval system SRS6 (EBI; http://srs.ebi.ac.uk) was used to access SWALL, IPI, and SWISS-PROT sequences. Gene Ontology assignments from the Gene Ontology Consortium (http://www.geneontology.org/) were accessed using the AMIGO gene ontology browser (http://www.godatabase.org/cgi-bin/go.cgi).

Identification of Cyclins and NIMA-Related Kinases (NEKs) Within RPS

Six hundred forty-two sequences containing the InterPro (Apweiler et al. 2001) motif cyclin (IPR004366) or cyclin c-term (IPR004367) were extracted from SWALL. For the NEKs no such motif exists; in this case, a keyword search was used on SWALL to extract sequences with a NIMA reference. The entries for these candidate NIMA sequences were then inspected manually to determine context, and 30 of these sequences were considered NIMA kinases. Literature searches were carried out to ensure that all known and homologous members were present in the bait sequence database.

In both cases the sequences were formatted as a database and then queried by the VPS. The sequences were ranked by e-value and length of alignment and evaluated on a case by case basis. This allowed us to identify known members, homologs, and novels.

Tree Analysis of Cyclins

Cyclin sequences were aligned using CLUSTAL V (Higgins et al. 1992). Alignments were edited in a text editor to trim the sequences down to the core cyclin domain, and then reloaded into CLUSTAL V. The sequences were then realigned and used to create a tree using the neighbor-joining method, with 1000 bootstraps. The tree was then visualized and printed using njplot (Perrière and Gouy 1996).

BLAST Searches

To identify homologous protein sequences, batch BLASTP searches were carried out using BLASTALL (http://www.ncbi.nih.gov/BLAST/). Results were parsed and, except where otherwise mentioned, an expectation value cutoff of e-30 was used to identify significant hits.

To map the cyclins and NEKs to ENSEMBL gene locations, the cDNA sequences corresponding to RPS and VPS sequences were extracted from RTS and compared with BLASTN against the ENSEMBL mouse cDNA sequences. The top three hits for each of the query sequences were examined manually to confirm the hit. This was useful for discriminating between cyclin B1 and its processed pseudogene, and for identifying Nek transcripts that bridged multiple ENSEMBL genes.

Gene Ontology-Based Bait Sequences (BAIT1)

Here, 1767 sequence identifiers were associated with cell cycle-related gene ontologies within AMIGO (http://www.godatabase.org/cgi-bin/go.cgi ; these were extracted and then used to query SWALL and SWISS-PROT using SRS6 (Etzold and Argos 1993; http://srs.ebi.ac.uk). Sequences corresponding to 1671 of the 1767 gene identifiers were found, sequences for 48 gene symbols could not be located, and 33 of these were the *Drosophila Scim* genes involved in female meiosis chromosome segregation. These represent mapped loci; however, no sequence is provided (Dobie et al. 2001). From these 1671 identifiers, 1859 sequences were extracted which formed the gene ontology-based bait sequence database (BAIT1).

Keyword Extraction of Cell-Cycle Sequences (BAIT2)

We used the following cell cycle-related keywords: cdk, cyclin, cell cycle, cdc, cell division, DNA damage, checkpoint, restriction point, mitotic, mitosis, cytokinesis, spindle, kinetochore, prophase, metaphase, anaphase, interphase, meiosis, meiotic, anaphase, prometaphase, telophase, SG2, G2M, MG1, S phase, M phase, G1 phase, and G2 phase, to extract 4437 eukaryotic FASTA sequences from SWALL and the IPI using SRS6. Queries were carried out with individual keywords, for example, *Query "([swall-AllText:mitosis])"*. Entries extracted from each of the individual searches were then merged into the BAIT2 sequence set. The evidence for each bait sequence, in this case, the associated keyword, was kept throughout the entire process to track the ontology suited to each entry. Careful manual curation assured that sequences where the keywords appeared out of context were not included in the BAIT2 sequence set.

Extraction of RPS Sequences Assigned Cell-Cycle Roles by MGI

Sequences from the RPS project were assigned GO ontologies by the MGI (Okazaki et al. 2002). Any sequence mapping to GO:0007049 or branches thereof were extracted in Microsoft Excel. Note that the MGI assignments were completed on the RPS and not the VPS. VPS is a superset of RPS; consequently, some assignments may have been missed by using the RPS.

ACKNOWLEDGMENTS

We thank the RIKEN Genome Exploration Research Group Phase I & II Team, Genomic Sciences Center, RIKEN, and the FANTOM consortium members. The Representative and Variable Protein sets used in these analyses were generated by the RTPS group, and the cell-cycle gene ontology assignments on RPS6 by the MGI Gene Ontology group. We also thank Dr. Rohan Teasdale for challenging discussions on the approach, and Professor David Hume, Institute for Molecular Bioscience, University of Queensland, without whom this collaboration would not have been possible.

REFERENCES

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains, and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Berke, J.D., Sgambato, V., Zhu, P.P., Lavoie, B., Vincent, M., Krause, M., and Hyman, S.E. 2001. Dopamine and glutamate induce distinct striatal splice forms of Ania-6, an RNA polymerase II-associated cyclin. *Neuron* **32:** 277–287.

Carnero, A. 2002. Targeting the cell cycle for cancer therapy. *Br. J. Cancer* **87:** 129–133.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

Dobie, K.W., Kennedy, C.D., Velasco, V.M., McGrath, T.L., Weko, J., Patterson, R.W., and Karpen, G.H. 2001. Identification of chromosome inheritance modifiers in *Drosophila melanogaster*. *Genetics* **157:** 1623–1637.

Etzold, T. and Argos, P. 1993. SRS—An indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.* **9:** 49–57.

Finley Jr., R.L. Thomas, B.J., Zipursky, S.L., and Brent, R. 1996. Isolation of *Drosophila* cyclin D, a protein expressed in the morphogenetic furrow before entry into S phase *Proc. Natl. Acad. Sci.* **93:** 3011–3015.

The Gene Ontology Consortium 2001. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11:** 1425–1433.

Groisman, I., Jung, M.Y., Sarkissian, M., Cao, Q., and Richter, J.D. 2002. Translational control of the embryonic cell cycle. *Cell* **109:** 473–483.

Ha Kim, Y., Yeol Choi, J., Jeong, Y., Wolgemuth, D.J., and Rhee, K. 2002. NEK2 localizes to multiple sites in mitotic cells, suggesting its involvement in multiple cellular functions during the cell cycle. *Biochem. Biophys. Res. Commun.***290:** 730–736.

Hames, R.S. and Fry, A.M. 2002. Alternative splice variants of the human centrosome kinase *Nek2* exhibit distinct patterns of expression in mitosis. *Biochem. J.* **361:** 77–85.

Higgins, D.G., Bleasby, A.J., and Fuchs, R. 1992. CLUSTAL V: Improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8:** 189–191.

Horton, L.E., Bushell, M., Barth-Baus, D., Tilleray, V.J., Clemens, M.J., and Hensold, J.O. 2002. p53 activation results in rapid dephosphorylation of the eIF4E-binding protein 4E-BP1, inhibition of ribosomal protein S6 kinase, and inhibition of translation initiation. *Oncogene* **21:** 5325–5334.

Kimura, S.H., Kataoka, T.R., Endo, Y., and Nojima, H. 1997. Genomic structure and chromosomal localization of mouse cyclin G1 gene. *Genomics* **46:** 483–486.

Kolonin, M.G. and Finley Jr., R.L. 2000. A role for cyclin J in the rapid nuclear division cycles of early *Drosophila* embryogenesis. *Dev. Biol.* **227:** 661–672.

Lock, L.F., Pines, J., Hunter, T., Gilbert, D.J., Gopalan, G., Jenkins, N.A., Copeland, N.G., and Donovan, P.J. 1992. A single cyclin A gene and multiple cyclin B1-related sequences are dispersed in the mouse genome. *Genomics* **13:** 415–424.

Lund, H., Eftedal, I., Haug, T., and Krokan, H.E. 1996. Pseudogenes for the human uracil-DNA glycosylase on chromosomes 14 and 16. *Biochem. Biophys. Res. Commun.* **224:** 265–270.

Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkoetter, M., Rudd, S., and Weil, B. 2002. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30:** 31–34.

Murray, A.W. and Marks, D. 2001. Can sequencing shed light on cell cycling? *Nature* **409:** 844–846.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Nigg, E.A. 2001. Mitotic kinases as regulators of cell division and its checkpoints. *Nat. Rev. Mol. Cell Biol.* **2:** 21–32.

Noguchi, K., Fukazawa, H., Murakami, Y., and Uehara, Y. 2002. *Nek11*, a new member of the NIMA family of kinases, involved in DNA replication and genotoxic stress responses. *J. Biol. Chem.* **277:** 39655–39665.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Osmani, S.A., May, G.S., and Morris, N.R. 1987. Regulation of the mRNA levels of nimA, a gene required for the G2-M transition in *Aspergillus nidulans*. *J. Cell Biol.* **104:** 1495–1504.

Perrière, G. and Gouy, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie* **78:** 364–369.

Peters, J.M. 2002. The anaphase-promoting complex: Proteolysis in mitosis and beyond. *Mol. Cell* **9:** 931–943.

Roig, J., Mikhailov, A., Belham, C., and Avruch, J. 2002. Nercc1, a mammalian NIMA-family kinase, binds the Ran GTPase and regulates mitotic progression. *Genes & Dev.* **16:** 1640–1658.

Rowley, R. 1992. Radiation-induced mitotic delay: A genetic characterization in the fission yeast. *Radiat. Res.* **132:** 144–152.

Sampath, D. and Plunkett, W. 2001. Design of new anticancer therapies targeting cell cycle checkpoint pathways. *Curr. Opin. Oncol.* **13:** 484–490.

Smits, V.A. and Medema, R.H. 2001. Checking out the G(2)/M transition. *Biochim. Biophys. Acta* **1519:** 1–12.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9:** 3273–3297.

Wassmann, K. and Benezra, R. 2001. Mitotic checkpoints: From yeast to cancer. *Curr. Biol.* **11:** 83–90.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13:** 1977–2000.

Xiong, Y., Menninger, J., Beach, D., and Ward D.C. 1992. Molecular cloning and chromosomal mapping of CCND genes encoding human D-type cyclins. *Genomics* **13:** 575–584.

Yaffe, M.B. 2002. How do 14-3-3 proteins work?—Gatekeeper phosphorylation and the molecular anvil hypothesis. *FEBS Lett.* **513:** 53–57.

WEB SITE REFERENCES

http://www.godatabase.org/cgi-bin/go.cgi; AMIGO gene ontology browser.

http://www.ncbi.nih.gov/BLAST; Basic Local Alignment Search Tool homepage.

http://cmap.nci.nih.gov/Ontologies; Cancer Analysis Molecular Project homepage.

http://www.ensembl.org/Mus_musculus; ENSEMBL mouse genome homepage.

http://fantom2.gsc.riken.go.jp; Functional Annotation of Mouse, RIKEN.

http://www.geneontology.org; Gene Ontology Consortium homepage.

http://www.ebi.ac.uk/IPI/IPI; International Protein Index.

http://www.ebi.ac.uk/interpro; InterPro homepage.

http://mips.gsf.de/proj/yeast/catalogues/; MIPS yeast catalogs.

http://www.informatics.jax.org/; MGI 2.8—Mouse Genome Informatics (MGI).

http://srs.ebi.ac.uk; Sequence Retrieval System (SRS6).

http://scop.mrc-lmb.cam.ac.uk/scop; Structural Classification of Proteins homepage.

Received December 3, 2002; accepted in revised form March 31, 2003.