Berkeley PHOG: PhyloFacts Orthology Group Prediction Web Server Supplement 2: Detailed Description of Experiments

Ruchira Datta[1], Bushra Samad[2], Christoph Neyer[2] and Kimmen Sjölander[1,2,3]

[1]QB3 Institute, University of California, Berkeley, [2] Department of Bioengineering, University of California, Berkeley, and [3] Department of Plant and Microbial Biology, University of California, Berkeley

Ortholog detection is essential in functional annotation of genomes, with applications to phylogenetic tree construction, prediction of protein-protein interaction and other bioinformatics tasks. The PHOG web server employs a novel algorithm to identify orthologs based on phylogenetic analysis. To assess the accuracy of PHOG, we used a set of human sequences and their predicted orthologs in three model organisms -- *Mus musculus* (mouse), *Danio rerio* (zebrafish) and *Drosophila melanogaster* (fruit fly) -- from the TreeFam-A resource as a gold standard benchmark. TreeFam-A uses a sophisticated ortholog-identification protocol (including tree reconciliation and manual curation) providing for a high-accuracy dataset. Mouse, zebrafish and fruit fly were selected since they had been targeted for analysis by both OrthoMCL-DB and InParanoid and represented a range of evolutionary distances.

Dataset selection: We chose a set of 100 human sequences from TreeFam-A meeting the following requirements:

1. Each sequence had to have orthologs in the TreeFam-A curated seed families from at least two of the three target species: mouse, zebrafish and *Drosophila melanogaster*.
2. No pair of sequences in the set could have a BLAST E-value < 1 (using a database size of 100,000) or share a common PFAM domain. This ensured that the dataset was filtered to remove potential homologs, so that method performance on this dataset should generalize.

We chose 100 human sequences as follows. We considered all human sequences for which TreeFam-A contained orthologs in at least two of the three target species: mouse, zebrafish, and fruitfly. This yielded a set of 353 human sequences. We then ran BLAST for each sequence in this set against the entire set, setting the database length to 100,000, and removed all sequences for which any other sequence in the set had a BLAST e-value <=1. This yielded a reduced set of 106 sequences. To further ensure that no pair of sequences were homologous, we searched for common PFAM domains by scoring each sequence against the PFAM-A library using hmmpfam (HMMs were downloaded from PFAM on January 13th, 2009). We found that four of the sequences had *no* significant PFAM hits (based on an e-value <= 0.001 and length >= 30); these four sequences were eliminated. Two pairs of sequences shared PFAM domains; we eliminated one member of each pair. This yielded a set of 100 sequences, containing a total of 138 PFAM domains. See Table S2-1.

Ensuring coverage: To ensure that each sequence in the test set was included in PhyloFacts, we identified the region of each sequence corresponding to each PFAM domain and clustered homologs and constructed a phylogenetic tree using the PhyloBuilder software (using global-local mode for domain-based homology and 5 subfamily HMM iterations). Users can submit their own sequences to this pipeline at the PhyloBuilder webserver at http://phylogenomics.berkeley.edu/phylobuilder.

Assessing method performance: For each human sequence and for each method, we identified predicted orthologs in mouse, zebrafish and *Drosophila melanogaster*, and compared these predicted orthologs against the orthologs identified by TreeFam-A.

We used TreeFam Release 6.0, InParanoid Version 6.1, OrthoMCL Version 2, and PHOG as of March 1st, 2009. For each of these methods, we used the precomputed predictions that were available on the respective websites. Ensembl Release 52 was used to cross-reference the identifiers used by some of the methods, along with the cross-reference provided by TreeFam. PHOG predictions were made on the basis of a total of 455 trees containing the human sequences based on trees in the PhyloFacts resource as of March 1st, 2009.

Cross-referencing database identifiers: In order to compare the ortholog predictions made by the various methods, we had to bring all the identifiers to the same form. The form we used was the Ensembl gene identifier. We used the gene_stable_id, transcript, transcript_stable_id, translation, and translation_stable_id tables provided by Ensembl to convert Ensembl transcript or protein identifiers to gene identifiers. We also used the stable_id_event table provided by Ensembl to convert outdated Ensembl identifiers to their present forms. We used the ens_xref table provided by TreeFam to convert UniProt accessions to Ensembl identifiers.

Statistical measures of performance: For each method, we computed the *recall* (or sensitivity) as the fraction of TreeFam-A orthology pairs found.

For reference, recall and precision are defined as follows:

Recall = TP/(TP+FN)

Precision = TP /(TP+FP)

where a *True Positive* (TP) is an orthology pair included in TreeFam-A that is also predicted by a method, a *False Negative* (FN) is an orthology pair included in TreeFam-A that is not predicted by a method (i.e., it is missed by the method), and a *False Positive* (FP) is an orthology pair predicted by a method that is not included in TreeFam-A.

The precision measure reported in the main body of the paper, reported as P1 in Supplement 3, is most stringent. This measure labels all predicted orthologs (by InParanoid, OrthoMCL or PHOG) not found by TreeFam-A as False Positives. In other words, if TreeFam did not identify an ortholog, and OrthoMCL did, OrthoMCL was wrong, and the computed precision will decrease.

We also computed a less stringent measure of precision, reported as P2 in Supplement 3, based on a different definition of False Positive. This measure only calls a predicted orthology pair an error in the case where TreeFam-A selected a *different* sequence from the same species. In other words, if for a particular human sequence TreeFam-A selects *no* ortholog from mouse, but OrthoMCL does, OrthoMCL might possibly be correct. That orthology pair does not contribute to the OrthoMCL recall, but does not hurt its precision. On the other hand, if TreeFam-A selects an ortholog from mouse (M1) but OrthoMCL selects a different ortholog from mouse (M2), the (H,M2) pair is called a false positive.

Results: We present the results in the Figures S2-1 through S2-4 below. The point labeled PHOG-T(M) on each curve describes the performance at threshold value 0.09375, where the precision and recall for the human-mouse ortholog predictions are approximately equal. The point labeled PHOG-T(Z) on each curve describes the performance at threshold value 0.296875, where the precision and recall for the human-zebrafish ortholog predictions are approximately equal. The point labeled PHOG-T(D) on each curve describes the performance at threshold value 0.9375, where the precision and recall for the human-fruitfly ortholog predictions are approximately equal. Detailed results are provided in Supplement 3.

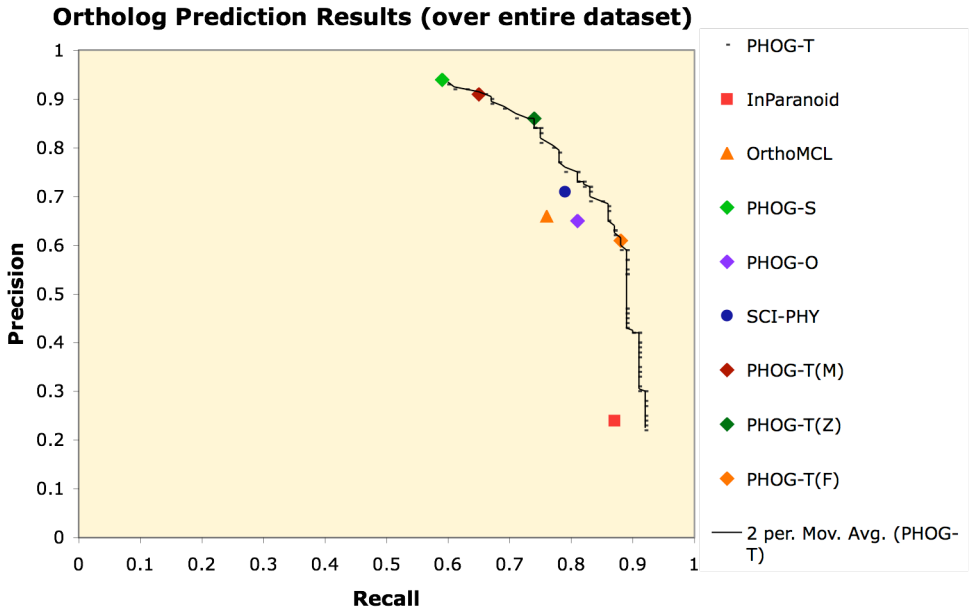

Figure S2-1. Precision-Recall results for all methods against the dataset as a whole.
Performance was evaluated on 100 human proteins selected from the TreeFam-A manually curated orthology database, with orthologs to each human protein from mouse, zebrafish and fruit fly. Methods evaluated include several PHOG variants, OrthoMCL-DB, InParanoid and SCI-PHY. PHOG-S represents super-orthology predictions, PHOG-O represents standard orthology predictions and PHOG-T represents the tree-distance thresholded variants. PHOG-T variants PHOG-T(M), -T(Z) and -T(F) correspond to tree-distance thresholds selected for optimal performance on this dataset for mouse, zebrafish and fruit fly respectively. Tree distance thresholds were 0.09375 (mouse), 0.296875 (zebrafish) and 0.9375 (fruit fly). SCI-PHY uses hierarchical clustering and encoding cost measures to define functional subtypes and is included for comparison. Recall measures the fraction of TreeFam-A orthologs detected by a method. Precision measures the fraction of a method's predicted orthologs that are included in TreeFam-A. A True Positive (TP) is an orthology pair included in TreeFam-A that is also predicted by a method, a False Positive (FP) is an orthology pair predicted by a method that is not included in TreeFam-A, and a False Negative (FN) is a TreeFam-A ortholog that is missed by a method.

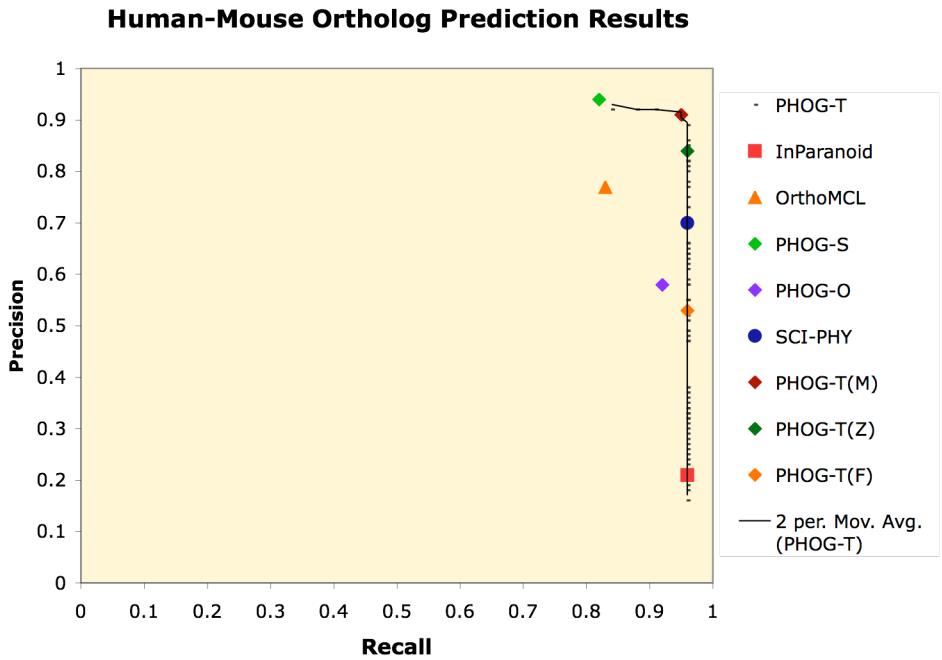

Figure S2-2. Results of ortholog prediction, restricted to human-mouse orthologs. See Figure S2-1 for details on precision and recall measures.

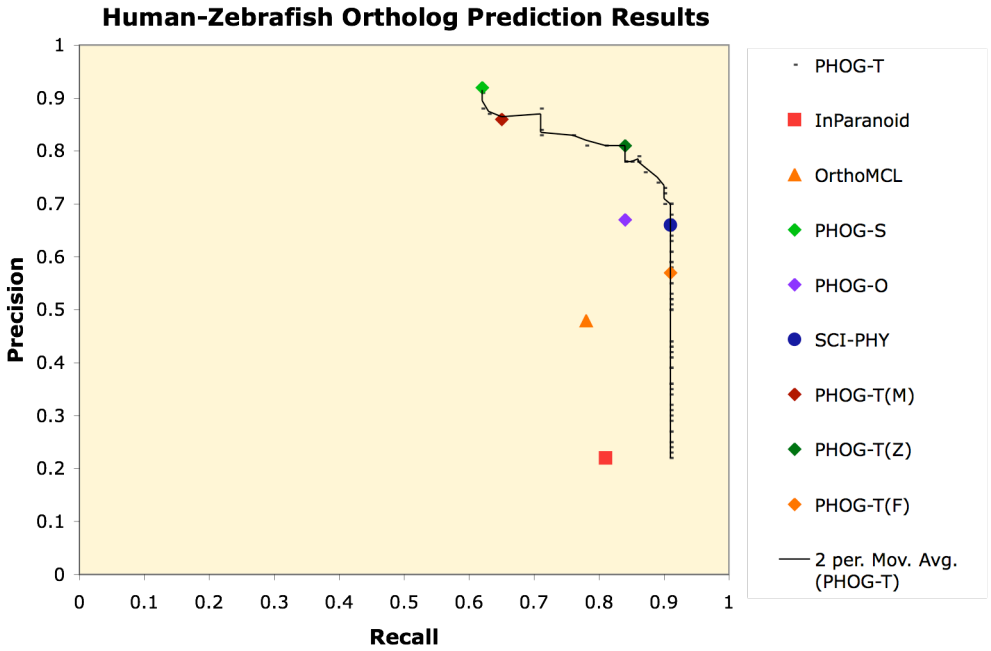

Figure S2-3. Results of ortholog prediction, restricted to human-zebrafish orthologs. See Figure S2-1 for details on precision and recall measures. OrthoMCL-DB's performance on zebrafish is uncharacteristically low, relative to its performance on other species.

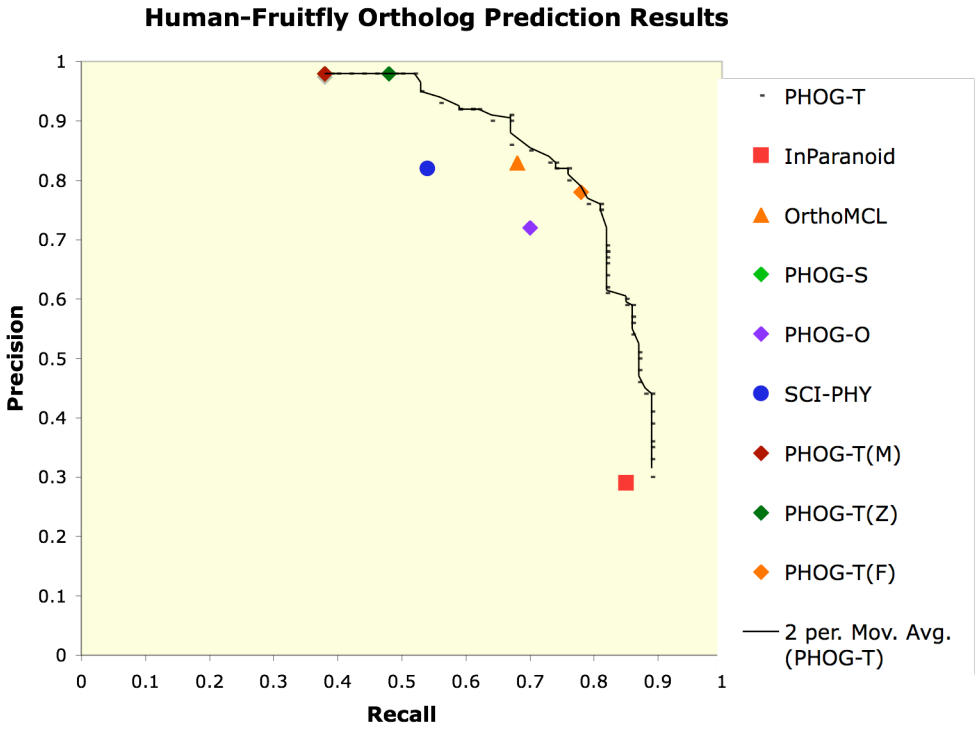

Figure S2-4. Results of ortholog prediction, restricted to human-fruit fly orthologs. See Figure S2-1 for details on precision and recall measures.

Table S2-1. Benchmark dataset of 100 human proteins taken from the TreeFam-A resource. The Ensembl identifier and equivalent UniProt accessions are shown, along with the description of each protein (obtained from the UniProt resource).

Ensembl ID	UniProt Accession	Description
ENSG00000065621	Q9H4Y5;Q5GM70;Q49TW5;Q86WP3	Glutathione S-transferase omega-2
ENSG00000168002	P52433;Q2M1Z4;P62487	DNA-directed RNA polymerase II subunit RPB7
ENSG00000163093	Q8N3I7;Q6PKN0	Bardet-Biedl syndrome 5 protein
ENSG00000137547	Q96Q54;Q9H0Y1;Q9P015	39S ribosomal protein L15
ENSG00000104980	O43615;Q8N193	Mitochondrial import inner membrane translocase subunit TIM44
ENSG00000105865	O95620;Q2NKK1	tRNA-dihydrouridine synthase 4-like
ENSG00000141101	Q7L6B7;Q7Z4B5;Q9NWB0;Q9ULX3;Q7M4M4	RNA-binding protein NOB1
ENSG00000198042	Q96SY6;Q9BXY0;Q5U5T1;Q86UC4	Protein MAK16 homolog
ENSG00000143486	P41214	Ligatin
ENSG00000154415	Q7KYM8;O43476;Q16821;A0AVQ2;Q86UI6;Q75LN8	Protein phosphatase 1 regulatory subunit 3A
ENSG00000102309	Q9Y237;Q5HYW6	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 4
ENSG00000172613	Q99638;Q96C41;Q6FI29	Cell cycle checkpoint control protein RAD9A
ENSG00000176876	Q9NR77	Peroxisomal membrane protein 2
ENSG00000196636	Q9NRP4;Q75MD6	Protein ACN9 homolog
ENSG00000100216	Q9NS69	Mitochondrial import receptor subunit TOM22 homolog
ENSG00000196683	O95939;Q9P0U1	Mitochondrial import receptor subunit TOM7 homolog
ENSG00000109519	Q549M6;Q9HAV7	GrpE protein homolog 1
ENSG00000147679	Q9BRU9;Q96NJ8	rRNA-processing protein UTP23 homolog
ENSG00000111639	Q9BQ36;Q96Q57;Q4U2R6;Q9P0N7	39S ribosomal protein L51
ENSG00000112096	Q16792;Q9P2Z3;Q96EE6;P78434;Q5TCM1;P04179	Superoxide dismutase
ENSG00000105197	Q96FJ5;Q6QA00;Q96GY2;Q9H370;Q3ZCQ8	Mitochondrial import inner membrane translocase subunit TIM50
ENSG00000167641	Q7Z4X7;Q96S54;Q96A00	Protein phosphatase 1 regulatory subunit 14A
ENSG00000139318	Q16828;Q53Y75;O75109;Q9BSH6	Dual specificity protein phosphatase 6
ENSG00000186184	Q5TBX2;Q9Y2S0;Q96BR3	DNA-directed RNA polymerases I and III subunit RPAC2
ENSG00000152137	Q9UKS3;Q6FIH3;Q9UJY1	Heat shock protein beta-8
ENSG00000145912	Q9NX24;Q9P095	H/ACA ribonucleoprotein complex subunit 2
ENSG00000198074	O60218;O75890	Aldo-keto reductase family 1 member B10
ENSG00000104320	Q63HR6;O60934;Q53FM6;Q7LDM2;Q32NF7;O60672	Nibrin
ENSG00000113569	Q9UFL5;Q9UBE9;O75694	Nuclear pore complex protein Nup155
ENSG00000093000	O75644;Q9UKX7;Q9NPR6;Q9NPM9;Q9P1K5	Nucleoporin 50 kDa
ENSG00000079246	Q0Z7V0;Q4VBQ5;P13010;Q53HH7;Q9UCQ1;Q9UCQ0;Q7M4N0;A8K3X5	ATP-dependent DNA helicase 2 subunit 2
ENSG00000134014	Q9BVF7;Q6AWB0;Q9H9T3;Q53G84;Q9NVZ1	Elongator complex protein 3
ENSG00000167799	Q6ZW59;Q8WV74	Nucleoside diphosphate-linked moiety X motif 8
ENSG00000148175	P27105;Q96FK4;Q15609;Q14087;Q5VX96	Erythrocyte band 7 integral membrane protein
ENSG00000125656	Q16740	Putative ATP-dependent Clp protease proteolytic subunit
ENSG00000088832	Q6LEU3;P62942;Q9H103;Q6FGD9;P20071;Q9H566;Q4VC47	Peptidyl-prolyl cis-trans isomerase FKBP1A
ENSG00000091651	Q9Y5N6	Origin recognition complex subunit 6
ENSG00000125037	Q9P0I2;Q53GH8;Q6ZMC2	Transmembrane protein 111
ENSG00000153037	P09132;Q96FG6	Signal recognition particle 19 kDa protein
ENSG00000159186	P48047;Q5U042;Q6IBI2	ATP synthase subunit O
ENSG00000089048	Q9H9Q5;Q9P1S6;Q9NX93;Q9H501;Q9HA35;Q86X92	ESF1 homolog
ENSG00000135046	P04083	Annexin A1
ENSG00000161217	P49585;Q86Y88;A9LYK9	Choline-phosphate cytidylyltransferase A
ENSG00000197265	Q9H2B9;P29084	Transcription initiation factor IIE subunit beta
ENSG00000176619	Q14734;O75292;Q96DF6;Q03252	Lamin-B2
ENSG00000116747	Q92787;P10155;Q9H1W6	60 kDa SS-A/Ro ribonucleoprotein
ENSG00000114388	Q9Y249;Q8WTW4;Q9Y497	Tumor suppressor candidate 4
ENSG00000121073	P78383;Q96EW7	Solute carrier family 35 member B1
ENSG00000112367	Q53H49;Q5TCS6;Q92562	SAC domain-containing protein 3
ENSG00000144231	Q52LT4;O15514	DNA-directed RNA polymerase II subunit RPB4

ENSG00000164258	Q9BS69;O43181	NADH dehydrogenase
ENSG00000149480	Q9UQB5;O94776	Metastasis-associated protein MTA2
ENSG00000111684	Q9BW40;Q6P1A2;Q92980;Q7KZS1	Lysophosphatidylcholine acyltransferase
ENSG00000178921	O15067	Phosphoribosylformylglycinamidine synthase
ENSG00000116903	Q5TE82;Q8IYI6	Exocyst complex component 8
ENSG00000099995	Q15459	Splicing factor 3 subunit 1
ENSG00000139180	Q2NKX0;Q14076;Q16795	NADH dehydrogenase
ENSG00000189091	Q15393;Q9UFX7;Q9UJ29;Q6NTI8;Q9BPY2;Q96GC0	Splicing factor 3B subunit 3
ENSG00000160325	Q8NBM6;Q9UGQ2;Q5SXD4	Transmembrane protein C9orf7
ENSG00000143815	Q59FE6;Q53GU7;Q14740;Q14739	Lamin-B receptor
ENSG00000089163	Q32M33;Q9Y6E7;O43346	NAD-dependent deacetylase sirtuin-4
ENSG00000116120	Q9NSD9;O95708;Q9NZZ6	Phenylalanyl-tRNA synthetase beta chain
ENSG00000128602	Q99835	Smoothened homolog
ENSG00000158042	Q9C066;Q6IAH8;Q9NRX2;Q96Q53	39S ribosomal protein L17
ENSG00000136197	Q9BPX7;Q9H779	UPF0415 protein C7orf25
ENSG00000132541	P52758;Q6FHU9	Ribonuclease UK114
ENSG00000150768	Q16783;Q53EP3;P10515	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
ENSG00000116729	Q7Z2Z9;Q5T9L3;Q8NC43	Integral membrane protein GPR177
ENSG00000162384	Q9NWV4	UPF0587 protein C1orf123
ENSG00000168496	P39748	Flap endonuclease 1
ENSG00000164219	Q5MJP9;P53609	Geranylgeranyl transferase type-1 subunit beta
ENSG00000138604	Q6GUQ2;O94923	D-glucuronyl C5-epimerase
ENSG00000129187	Q9BVD8;P32321	Deoxycytidylate deaminase
ENSG00000113460	Q96DH1;Q3ZTT4;Q8TDN6;Q8N453	Brix domain-containing protein 2
ENSG00000067533	Q9Y3B9	RRP15-like protein
ENSG00000180875	Q86UD9;Q9H772	Gremlin-2
ENSG00000125207	Q8NA60;O95404;Q96JD5;Q8TBY5;Q96J94	Piwi-like protein 1
ENSG00000146670	Q96FF9	Sororin
ENSG00000164109	Q13257	Mitotic spindle assembly checkpoint protein MAD2A
ENSG00000070785	Q5QP90;Q5QP89;Q9H850;Q8NDB5;Q8WV57;Q9NR50	Translation initiation factor eIF-2B subunit gamma
ENSG00000169021	P47985;Q9UPH2;Q6NVX5	Cytochrome b-c1 complex subunit Rieske
ENSG00000105141	O95823;Q3SYC9;P31944	Caspase-14
ENSG00000196510	Q9BU24;Q96GF4;Q9NT16;Q96AC4;Q9UJX3	Anaphase-promoting complex subunit 7
ENSG00000124383	O00566;A0AVJ8	U3 small nucleolar ribonucleoprotein protein MPP10
ENSG00000100307	O95931;Q86T17	Chromobox protein homolog 7
ENSG00000063660	P35052	Glypican-1
ENSG00000167797	O75956	Cyclin-dependent kinase 2-associated protein 2
ENSG00000099341	P48556	26S proteasome non-ATPase regulatory subunit 8
ENSG00000113575	P05323;P67775;P13197	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform
ENSG00000134153	Q96ED5;Q9NPA0	UPF0480 protein C15orf24
ENSG00000103035	Q96E97;P51665;Q6PKI2	26S proteasome non-ATPase regulatory subunit 7
ENSG00000151247	Q96E95;P06730	Eukaryotic translation initiation factor 4E
ENSG00000103245	Q9H6Q4;Q96S10;Q9H6J8;Q53GC6;A1L385	Nuclear prelamin A recognition factor-like protein
ENSG00000138495	Q14061;Q3MHD6	Cytochrome c oxidase copper chaperone
ENSG00000165271	Q9H6R4	Nucleolar protein 6
ENSG00000183093	Q9HCA2;Q9UQJ3;Q969U0;P52815	39S ribosomal protein L12
ENSG00000047315	P30876;Q8IZ61	DNA-directed RNA polymerase II subunit RPB2
ENSG00000104472	Q9NRG0	Chromatin accessibility complex protein 1
ENSG00000153879	Q5U052;P53567	CCAAT/enhancer-binding protein gamma
ENSG00000012061	P07992	DNA excision repair protein ERCC-1