**Supplementary Material**

# Computational characterization of multiple Gag-like human proteins

**Mónica Campillos[1], Tobias Doerks[1], Parantu K. Shah[1,2] and Peer Bork[1]**

[1]EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]Present address: Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA
*Corresponding author:* Bork, P. (bork@embl.de) and Campillos, M. (campillo@embl.de).

**In a genome-wide analysis, we have identified 85 human genes encoding 103 protein isoforms that resemble retroviral Gag proteins. These genes were domesticated from retrotransposons in at least five independent events during vertebrate evolution and were subsequently duplicated further in mammals. Structural insights into the mammalian proteins can be inferred by homology to Gag from viruses such as HIV; in turn, the cellular roles of the mammalian Gag homologs, such as apoptosis-related functions and binding to ubiquitin ligases, might hint at further functionality of viral Gag itself.**

Published in *Trends in Genetics* (2006) Vol. 22, issue 11

The Supplementary Material is organized in the following sections:

# 1. METHODS

## a) Detection and prediction of the Gag-like domesticated proteins

### Genome wide search

Hidden Markov models (HMM) of the N-terminal capsid domains were created by seeding the Retrotransposon Gag domain[1] (PF03732) with a translated Repbase database (version 10.2)[2]. The HMM profile created using HMMER package (http://hmmer.wustl.edu) was applied to all six open reading frames of the human genome (NCBI Build 35) (Table S2.1). In order to detect the most likely ancestral retrotransposon of Gag-like human genes the human chromosomal hits obtained were compared to sequences in the Repbase[2] using tblastx[3] (TableS2.2).

For a comprehensive identification of Gag-like sequences in seven vertebrate genomes (see below) we performed PSI-BLAST searches (e-value < $10^{-10}$) with position specific matrices (created using the translated Repbase) of ancestral Gag of retroransposons. This fast computational approach allowed the detection of domain independent acquisitions because it compares the whole retrotransposon sequence with DNA sequences. A PSI-BLAST search with GYPSYDR1 retrieved proteins with Scan domains instead of the SASPase protein (chr2 hit in Table S2.1) (Table S2.2).

In order to find retrotransposon sequences more similar to SASPase than GYPSYDR1 we searched the human SASPase sequence against vertebrate genomes using blast and a reconstructed part of a retrotransposon in *Danio Rerio* genome ("predicted" in Figure1). We then performed PSI-BLAST searches in vertebrate genomes with the "predicted" retrotransposon (Table S2.2).

The vertebrates genomes searched were: *Homo sapiens* (NCBI Build 35), *Mus musculus* (NCBI Build 34), *Rattus novergicus* (version 3.1 Atlas group), *Gallus gallus* (produced by the Genome Sequencing Center at the Washington University School, February 2004), *Xenopus tropicalis* (version 3.0 DOE Joint Genome Institute (JGI)), *Danio rerio*, (Zv4 assembly) and *Takifugu rubripes* (v.3.0, JGI).

### Protein prediction

The Gag-like proteins encoded in the vertebrate genomes were identified using tblastn[3] against the non-redundant database (NR) at NCBI (May 2005), Ensembl protein predictions (march 2005) and retrotransposon sequences annotated in Repbase database. Gag-like proteins without a tblastn hit higher than 98% identity were predicted using genewise[4] and both strands of chromosomal region of the Gag-hit (extended 2kb in both 5' and 3' directions).

## b) ESTs and orthology detection

For the assessment of EST support ESTs from dbEST [5] (January 2005) were aligned against the Gag-like human genes using blastn[3] and we consider only EST alignments with a percentage identity greater than 96% and the alignment length greater than 100 bases. Orthology was determined using phylogenetic analysis and manual synteny determination. Mouse and rat orthology for human genes with scan domains was directly obtained from Ensembl database. ESTs specificity is not possible to determine for Hs_PMNA6A and Hs_PMNA6B because their sequences are 100% identical.

**c) Phylogenetic analysis**

For the phylogenetic analysis shown in Figure 1, a neighbour-joining tree was prepared after aligning representative reverse transcriptase sequences (~240 aa) of the different LTR-retroelements clades following a similar previous analysis [6]. The tree root is a non LTR-retrotransposon element (not shown). Nodes were collapsed in all clades with the exception of Ty3/Gypsy group. The *Danio rerio* and *Takifugu rubripes* retrotransposons related to the domesticated mammalian proteins are marked in bold. The reverse transcriptase sequences of the "*predicted*" retrotransposon were predicted using the protein alignment of the reverse transcriptase sequences of LTR retrotransposons and the translated DNA region in the *Danio Rerio* genome (chr17 33584814-33585179) similar to the SASPase human sequence.

The proteins used to build the tree are the following: **Retroviruses,** Human immunodeficiency virus type 1 (HIV1) P03367, Rous sarcoma virus (RSV) P03354, Human T-cell leukemia virus typeI (HTLV1) P03362, Human immunodeficiency virus type 2 (HIV2) P04584, Mouse intracisternal a-particle (IAP) P11368, Feline immunodefiency virus (FIV) P16088, Bovine leukemia virus (BLV-1) P03361, Feline leukemia virus (FeLV) AAA93092, HERV-K Human NP_001007237, Foamy Virus-Human NP_044450, Mouse mammary tumor virus (MMLV) NP_955591. Caulimoviruses: Cauliflower mosaic virus (CAMV) NC_001497, Commelina yellow mottle virus (CoYMV) NP_039820, Petunia vein clearing virus (PVCV) NC_001839. **DIRS**: DIRS-like-D.Rerio *Danio Rerio* DIRS1_DR Repbase, DIRS1-Dic *Dictyostelium Discoideum* M11339, DIRS-like-S.Purpuratus *Strongylocentrotus Purpuratus* retrobase http://biocadmin.otago.ac.nz/retrobase/, DIRS-like-T.nigroviridis *Tetraodon nigroviridis* AF442732. **Hepadnaviruses**: HBV-Duck NC_001344, HBV-Human NP_647604, HBV-Orangutan AF193864, HBV-Woodchuck. **BEL**: TAS *Ascaris lumbricoides* Z29712, BEL *Drosophila melanogaster* U23420, Cer11 *Caenorhabditis elegans* AAA82437, Ninja *Drosophila simulans* T31674, Pao *Bombyx mori* L09635. **Ty1/COPIA**: PAT *Panagrellusredivivus* X60774, SIRE1 *Glycine max* AF053008, 1731 *Drosophila melanogaster* X07656, Evelknievel *Arabidopsis thaliana* AF039373, Copia *Drosophila melanogaster* X04456, Tnt1 *Tobacco* X13777, BARE1 *Hordeum vulgare* Z17327, Ty5 S.paradoxus U19263, Tca5 Candida albicans AF065434, Osser CAA49283 *Volvox carteri f. nagariensis*, Ty1 *Saccharomyces cerevisiae* M18706, Ty4 *Saccharomyces cerevisiae* M94164. Reverse transcriptase sequences of the **Ty3/Gypsy** lineage were taken from the alignment of this sequences[7] with the exception of the Saci-2 *Schistosoma mansoni* protein DAA04499.

**Evolutionary trees for Sushi, Paraneoplastic and Scan families**

Vertebrate proteins of each family and retrotransposons from Repbase[2] were aligned using probcons[8] and refined manually by seaview[9]. Fragments and identical sequences were discarded from the alignment. Conserved blocks of aligned sequences were selected with Gblocks[10]. Phylogenetic analyses were performed with MRBAYES version 3.0[11] (4 Markov chains, $10^6$ generations, with each 20$^{th}$ tree sample). MRBAYES was run with average rate matrix for amino acid data, general time reversible (GTR) substitution model with among site rate heterogeneity following a gamma invariant sites distribution. Bayesian posterior probabilities were estimated on the consensus of the last 5000 trees using Clann[12]. The same topology was obtained by using the maximum likelihood (ML) method as implemented in PHYML[13] (under the Jones–Taylor–Thornton (JTT) and fixed proportion of invariable sites and one category of substitution rate) and neighbour-joining method (NJ) (as above). Trees were rooted with Gag proteins from retrotransposon sequences in Repbase.

**d) Structural alignment of N- and C-_terminal domains of Gag capsid protein.**

**Method for the p-value calculation**

Murzin proposed a p-value based on the binomial distribution to suggest the likelihood that a sequence identity found after structure-based alignment could occur by chance[14]. Given *n* conserved sites between two similar protein structures, he suggested that the probability that *m* of these sites would contain identical amino acids would be:

$$P(m) = \left[\frac{n!}{m!(n-m)!}\right]\bar{p}^{m}(1-\bar{p})^{n-m} \quad \text{provided } m > m_o + \sigma$$

where $\bar{p}$ is the mean probability of finding identical residues at structurally equivalent sites, $m_o = n\bar{p}$ (where binomial has maximum) and $\sigma = \sqrt{n\bar{p}(1-\bar{p})}$ (the half-width of approximating distribution). Murzin suggested that the value of $\bar{p}$ for similarly folded structures is about 1/15 and almost certainly is less than 1/10. This calculation was originally applied to the cystatin-monellin similarity, where an evolutionary relationship was inferred based on a p-value ~ $10^{-3}$ [14].

**Assessing statistical significance of the alignment of duplicated regions**

We used combinatorial extension (CE) web server (http://cl.sdsc.edu/ce/ce_align.html) to prepare initial structure based alignment between N and C-terminal capsid domain of HIV Gag protein[15]. We used high resolution crystal structures deposited in PDB with ids 1M9X (chain H containing the N-terminal region) and 2BUO (chain A containing the C-terminal region) for the alignment. The alignment suggested by CE contained first four out of five helices the N-terminal region and all four helices from the C-terminal region. The alignment provided by CE was given as an input to STAMP program [16]. STAMP aligned 30 out of total 49 Cα positions in the alignment with a root of mean square deviations (RMSD) of 2.11 Å, which is suggestive of a significant structural similarity. This alignment was modified manually taking into consideration structural equivalences and information from multiple sequence alignment. The final alignment had 11 identities and 5 conservative substitutions in 55 aligned positions on which the p-value calculation was applied.

.

$P(m) = 6.64*10^{-4}$ ($\bar{p}$ =1/15, $n$ = 55 , $m$ = 11 ) and

$P(m) = 1.11*10^{-3}$ ($\bar{p}$ = 1/10, $n$=55, $m$ = 11)

Conservative substitutions at aligned positions perform the same function Therefore, taking conservative substitutions in account the p-values are even more significant

P(m)= 3.072 *10-7 ($\bar{p}$ = 1/15, $n$ = 55, $m$ = 16) and

P(m)= 4.8 *10-5 ($\bar{p}$ =1/10, $n$=55, $m$ =16).

### e) Identification of Gag-like proteins interaction partners

Protein interaction partners of mammalian (*Homo sapiens*, *Mus musculus, Rattus novergicus*) Gag-like proteins were extracted from the NCBI file interactions.gz, which contains interactions extracted form literature from the BIND[17] (Biomolecular Interaction Network) and HPRD[18] (Human Protein Reference Database) databases. Additional interactions described in literature were added manually. Proteins were classified manually in different categories. We determined statistically enriched GO terms in the mammalian Gag-like proteins and their interaction partners (hypergeometric test, $p<0.05$, Benjamini&Hochberg false Discovery Rate correction) by comparing the GO terms associated to these proteins to the terms associated to all mammalian genes using BINGO[19].

## 2. GENOME WIDE SEARCHES FOR GAG-LIKE PROTEINS; PROTEIN DETAILS

### a) Genome wide searches for Gag-retrotransposon like proteins

| HMM HIT | REPBASE HIT | E-value tblastx |
|---|---|---|
| chr14_73249245..73249673 | Gypsy9-I_DR | 2.E-37 |
| chrX_151829866..151830294 | Gypsy12-I_DR | 3.E-27 |
| chr14_100419905..100420348 | RONIN2_I | 1.E-22 |
| chrX_117741163..117741591 | Gypsy12-I_DR | 2.E-22 |
| chr14_92719478..92719906 | Gypsy12-I_DR | 6.E-26 |
| chr8_26421326..26421754 | Gypsy12-I_DR | 3.E-32 |
| chrX_103165021..103165449 | Gypsy12-I_DR | 3.E-22 |
| chr22_43213303..43213746 | RONIN2_I | 4.E-14 |
| chrX_71133588..71134031 | RONIN2_I | 6.E-17 |
| chrX_139996200..139996643 | RONIN2_I | 3.E-11 |
| chrX_151896412..151896840 | Gypsy9-I_DR | 7.E-27 |
| chr7_93937689..93938132 | RONIN2_I | 1.E-23 |
| chrX_133881570..133881998 | RONIN2_I | 8.E-12 |
| chrX_133911219..133911647 | RONIN2_I | 3.E-11 |
| chrX_133891944..133892372 | RONIN2_I | 3.E-11 |
| chrX_152069293..152069724 | Gypsy9-I_DR | 2.E-05 |
| chrX_151915085..151915516 | Gypsy12-I_DR | 2.E-14 |
| chrX_151912418..151912849 | Gypsy12-I_DR | 2.E-14 |
| chrX_152183821..152184252 | Gypsy10-I_DR | 3.E-16 |
| chr8_143691670..143692092 | Gypsy-26-I_DR | 3.E-06 |
| chr22_18213662..18214105 | SUSHIIDR1 | 6.E-05 |
| *chr2_70099813..70100241 | GYPSYDR1 | 3.E-04 |
| chrX_151868476..151868928 | Gypsy10-I_DR | 1.E-07 |
| chrX_77718975..77719418 | RONIN2_I | 4.E-11 |

**Table S2.1.** *Homo sapiens* chromosome hits of N-terminal Capsid profile, the best Repbase element tblastx and the E-value of the comparison. *This hit was not retrieved using PSI-BLAST searches with GYPSYDR1 retrotransposon.

| *RETROTRANSPOSON* | *RONIN2_I* | *Gypsy12-I_DR* | *Gypsy-26-I_DR* | *GYPSYDR1* | *Predicted* |
|---|---|---|---|---|---|
| **DOMESTICATED PROTEINS** | **Sushi family** | **Paraneoplastic family** | **ARC** | **SCAN family** | **SASPase** |
| ***Homo Sapiens*** | 12 | 16 | 1 | 6 | 1 |
| ***Mus Musculus*** | 12 | 12 | 1 | 3 | 1 |
| ***Rattus Novergicus*** | 7 | 13 | 1 | 2 | 1 |
| ***Gallus Gallus*** | 0 | 0 | 1 | 0 | 0 |
| ***Xenopus Tropicalis*** | 743 | 228 | 1 | 9 | 10 |
| ***Takifugu ruprides*** | 342 | 16 | 0 | 0 | 0 |
| ***Danio Rerio*** | 1482 | 210 | 161 | 645 | 4 |

**Table S2.2.** Number of hits in vertebrate genomes obtained by PSI-BLAST searches with each representative retrotransposon Gag protein.

**b) Protein IDs and chromosomal locations of Gag-like proteins**

**Table S2.3 (in a separate Excel file)** Chromosomal positions and protein identifiers of domesticated Gag and retrotransposon sequences. "pred" denotes that the protein was predicted in this work.

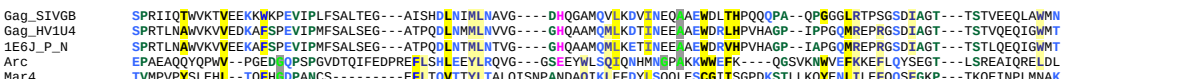
**c) ESTs and rat and mouse orthology for human Gag-like proteins**

**Table S2.3 (in a separate Excel file) Human_genes_ESTs** Gag-like human genes, ESTs associated to them and presence of orthologous genes in Mouse and Rat genomes.

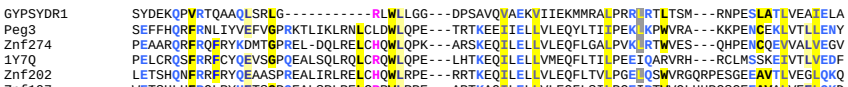
# 3. ALIGNMENTS OF DOMAINS IN Gag PROTEINS

## MATRIX DOMAIN ALIGNMENT

## N-TERMINAL CAPSID DOMAIN ALIGNMENT



## C-TERMINAL CAPSID DOMAIN ALIGNMENT



**FigureS3.** Alignments of the Matrix, N-terminal and C-terminal capsid domains of representative Gag-like human proteins, retrotransposons and Human and simian immunodeficiency viruses. The alignment is coloured by chroma[20] (conserved prolines: white on grey; conserved glycines and alanines: green on grey; conserved leucines, isoleucines, phenylalanines, cysteines, valines and tyrosines: yellow on grey; conserved asparagines and glutamines: dark red on grey; conserved glutamic acids: light red on grey; conserved threonines and serines: light blue on grey; conserved aliphatic residues: grey on yellow; conserved hydrophobic residues: black on yellow; conserved small residues: dark green on white; conserved positively charged residues: blue on white; conserved polar residues: dark blue on white; conserved charged residues: pink on white; conserved aromatic residues: blue on yellow; conserved big residues: blue on light yellow; conserved negatively charged residues: red on white).

The consensus sequence (conserved in 80% of the sequences) shown below; h, p, s, l, b, c, a, + and - indicate hydrophobic, polar, small, aliphatic, big, charged, aromatic, positively charged and negatively charged residues, respectively.

## 4. N- AND C-_TERMINAL DOMAINS OF Gag CAPSID PROTEIN ARE DUPLICATED DOMAINS

In order to provide support to our hypothesis that there has been an ancient gene duplication in HIV Gag protein; we aligned the N and C-terminal fragments of the protein using structure based sequence alignment methods and applied the method of Murzin[14] (see Methods above) to assign statistical significance to sequence identities found after structural alignment.

Both the RMSD and the p-value (see Methods above) suggest that observed structural and sequence similarities are statistically significant and provide basis for the gene duplication event.

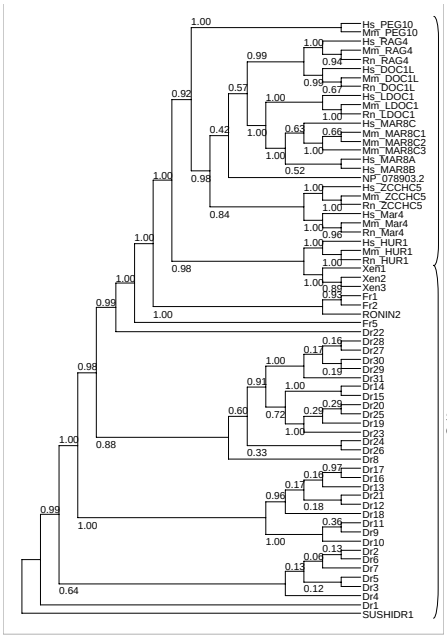
```
1E6J_P_N          SPRTLNAWVKVVEEKAFSPEVIPMFSA-LS------EGATPQ-DLNTMLNTVGGHQAAMQ
E6J_P             T-SILRQGPKEP----F-RDYVDRFYKTLRAEQASQEVKNWMTE--TLLVQNA-NPDCKT
Consensus/80%     *.phLp.hsK.s....F..-hlsbF...Lp......Es.s.b.-..TbLspst.p.ssbp

1E6J_P_N          MLKETINEEAAE
E6J_P             ILKALGPAATLE
Consensus/80%     bLK.hhs..shE
```

**FigureS4. Alignment of the N-terminal and C-terminal capsid domains of HIV-1.** See Figure S3 for color code.

## 5. EVOLUTIONARY ASPECTS

We created phylogenetic trees of vertebrate proteins from the five families (we show here only Sushi, Paraneoplastic and Scan families) in order to determine the evolutionary history of the domestication events. The phylogenetic analysis suggested that the domestication occurred once for the Paraneoplastic and Scan family. The domesticated mammalian Sushi proteins are more likely the result of two independent domestication events, one would result in the HUR1 proteins and the other in the rest of Sushi proteins. The domestication events of these three families and SASPase family occurred before the *Metatherian-Eutherian* split while the Arc family was domesticated before the divergence of amphibian.

Chromosomal positions of vertebrate sequences of each family are shown in TableS2_3



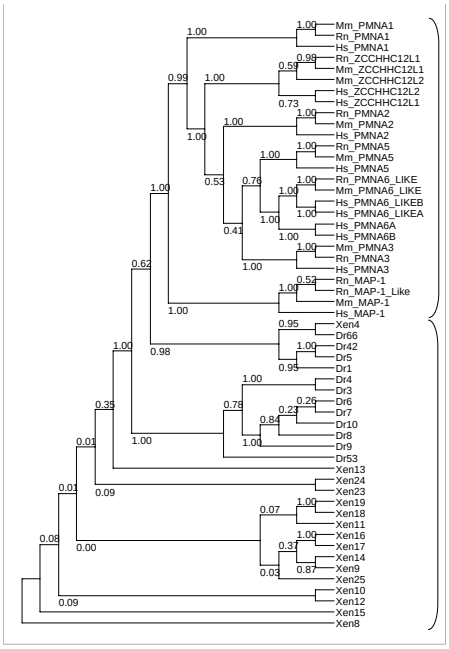




**Figure S5.1 (left).** Phylogenetic tree of Mammalian Gag-like proteins originated from Sushi-like retrotransposons (only posterior probabilities > 0.5 are shown). The alignment is 74 amino-acid long and contains sequences of the N-terminal Capsid domain. Mammalian RAG1 proteins were discarded with the aim of increasing the sequence information (they only retain half of the domain).

**Figure S5.2 (right).** Phylogenetic tree of Mammalian Gag-like proteins originated from Gypsy12-like retrotransposons (only posterior probabilities > 0.5 are shown). The alignment is 142 amino-acid long and cover sequences of the N-terminal and C-terminal Capsid domains. Mammalian proteins that retain only the Matrix domain were not included in this analysis to increase the sequence information of the alignment.

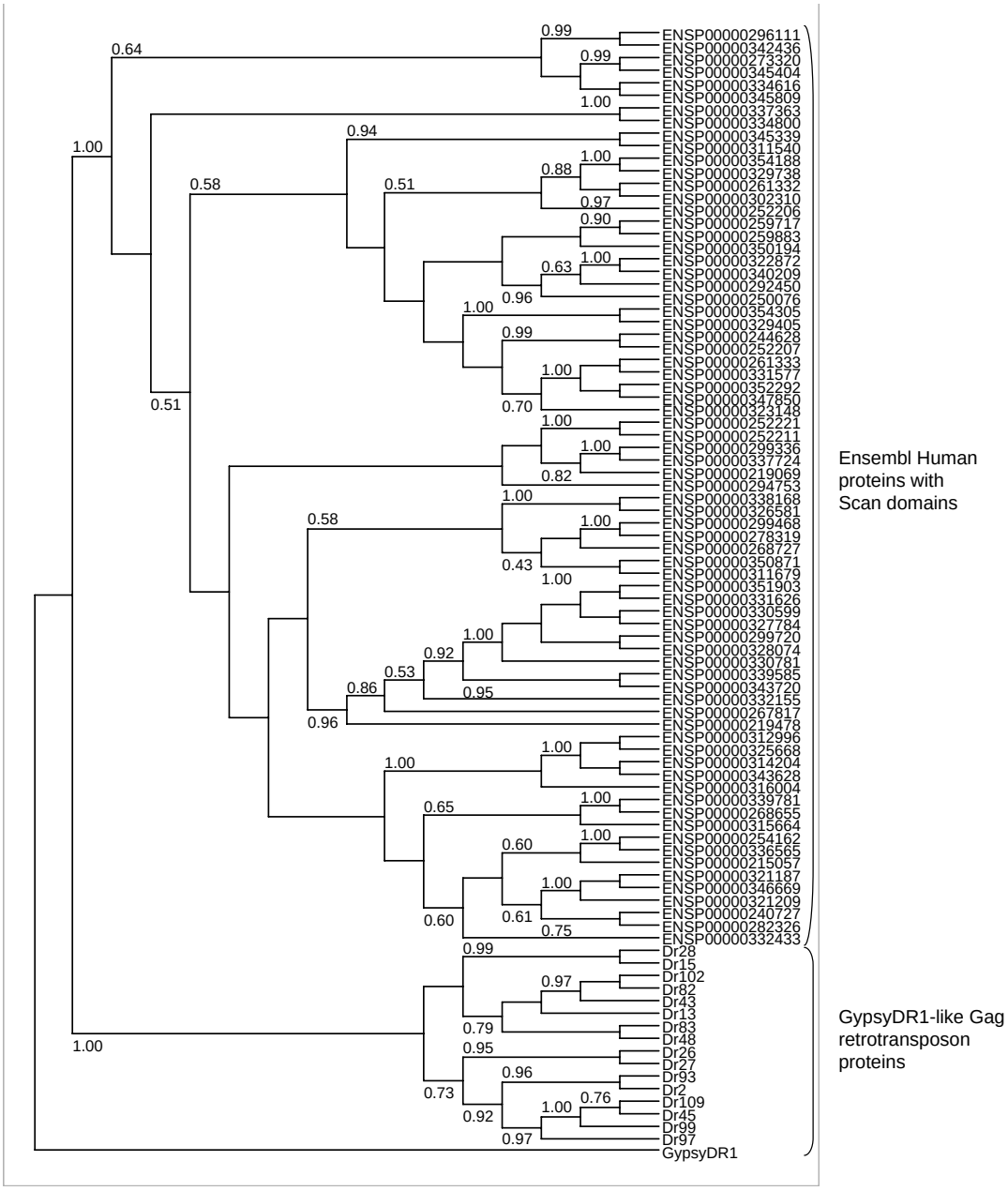



**Figure S5.3.** Phylogenetic tree of *Homo sapiens*, *Xenopus tropicalis* and *Danio rerio* proteins with scan domain originated from GypsyDR1-like retrotransposons (only posterior probabilities > 0.5 are shown). Human sequences with scan domain were obtained from Ensembl database; only proteins with a complete scan domain were analyzed. The alignment is 81 amino acids long.

# 6. INTERACTION PARTNERS FOR GAG-LIKE PROTEINS

**Table 6.1** Gag-like protein interaction parterns

| Gag-LIKE PROTEIN | INTERACTION PARTNER | REF. |
|---|---|---|
| ZNF496 | NSD1 | [21] |
| Zfp110 | TRAF6 | [22] |
| Zfp110 | NGFR | [22] |
| Scand1 | Pparg | [23] |
| ZNF24 | ZNF174 | [24] |
| PEG3 | SIAH1 | [25] |
| PEG3 | TRAF2 | [26] |
| PEG3 | SIAH2 | [25] |
| ZNF197 | VHL | [27] |
| ZNF396 | ZNF397 | [28] |
| SCAND1 | ZNF42 | [29] |
| SCAND1 | ZNF202 | [30,31] |
| ZNF42 | LDOC1 | [32] |
| ZNF174 | ZNF20 | [24] |
| MAP-1 | RASSF1 | [33] |
| MAP-1 | TNFRSF10A (TNR10A) | [33] |
| MAP-1 | BCL2 | [34] |
| MAP-1 | BCL2L1 | [34] |
| MAP-1 | TNFRSF1A (TNR1A) | [33] |
| MAP-1 | BAX | [34, 15, 35] |
| PEG10 | SIAH1 | [35] |
| PEG10 | ACVRL1 | [36] |
| PEG10 | SIAH2 | [35] |
| LDOC1 | CEBPE | [37] |
| ARC | MAP2 | [38] |
| ARC | Actin-F | [39] |
| ARC | CaMKII | [40] |

**Table 6.2.** Selected GO terms statistically enriched in mammalian Gag-like proteins and their interaction partners

| BIOLOGICAL ACTIVITY (GO terms) | P value |
|---|---|
| Apoptosis (GO:6915) | $1.2.10^{-7}$ |
| Regulation of transcription, DNA-dependent (GO:6355) | $1.2.10^{-5}$ |
| Protein binding (GO:45308) | $2.7.10^{-4}$ |
| Protein ubiquitination (GO:16567) | $3.5.10^{-4}$ |
| Anti-apoptosis (GO:6916) | $4.2.10^{-4}$ |
| Cell cycle (GO:7049) | $5.8.10^{-4}$ |
| Ubiquitin ligase complex (GO:151) | $3.3.10^{-3}$ |

## REFERENCES


1. Nakayashiki, H. et al. Pyret, a Ty3/Gypsy retrotransposon in Magnaporthe grisea contains an extra domain between the nucleocapsid and protease domains. *Nucleic Acids Res* **29**, 4106-13 (2001).
2. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7 (2005).
3. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**, 2444-8 (1988).
4. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-95 (2004).
5. Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. dbEST--database for "expressed sequence tags". *Nat Genet* **4**, 332-3 (1993).
6. Malik, H.S. & Eickbush, T.H. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11**, 1187-97 (2001).
7. Malik, H.S. & Eickbush, T.H. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J Virol* **73**, 5186-90 (1999).
8. Do, C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **15**, 330-40 (2005).
9. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**, 543-8 (1996).
10. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
11. Ronquist, F. & Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-4 (2003).
12. Creevey, C.J. & McInerney, J.O. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**, 390-2 (2005).
13. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
14. Murzin, A.G. Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors. *J Mol Biol* **230**, 689-94 (1993).
15. Shindyalov, I.N. & Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-47 (1998).
16. Russell, R.B. & Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14**, 309-23 (1992).
17. Alfarano, C. et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**, D418-24 (2005).
18. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-71 (2003).
19. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-9 (2005).
20. Goodstadt, L. & Ponting, C.P. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics* **17**, 845-6 (2001).
21. Nielsen, A.L. et al. Nizp1, a novel multitype zinc finger protein that interacts with the NSD1 histone lysine methyltransferase through a unique C2HR motif. *Mol Cell Biol* **24**, 5184-96 (2004).
22. Gentry, J.J., Rutkoski, N.J., Burke, T.L. & Carter, B.D. A functional interaction between the p75 neurotrophin receptor interacting factors, TRAF6 and NRIF. *J Biol Chem* **279**, 16646-56 (2004).
23. Castillo, G. et al. An adipogenic cofactor bound by the differentiation domain of PPARgamma. *Embo J* **18**, 3676-87 (1999).
24. Williams, A.J., Blacklow, S.C. & Collins, T. The zinc finger-associated SCAN box is a conserved oligomerization domain. *Mol Cell Biol* **19**, 8526-35 (1999).
25. Relaix, F. et al. Pw1/Peg3 is a potential cell death mediator and cooperates with Siah1a in p53-mediated apoptosis. *Proc Natl Acad Sci U S A* **97**, 2105-10 (2000).
26. Relaix, F., Wei, X.J., Wu, X. & Sassoon, D.A. Peg3/Pw1 is an imprinted gene involved in the TNF-NFkappaB signal transduction pathway. *Nat Genet* **18**, 287-91 (1998).
27. Li, Z. et al. The VHL protein recruits a novel KRAB-A domain protein to repress HIF-1alpha transcriptional activity. *Embo J* **22**, 1857-67 (2003).
28. Wu, Y. et al. Identification and characterization of two novel human SCAN domain-containing zinc finger genes ZNF396 and ZNF397. *Gene* **310**, 193-201 (2003).
29. Sander, T.L., Haas, A.L., Peterson, M.J. & Morris, J.F. Identification of a novel SCAN box-related protein that interacts with MZF1B. The leucine-rich SCAN box mediates hetero- and homoprotein associations. *J Biol Chem* **275**, 12857-67 (2000).
30. Schumacher, C. et al. The SCAN domain mediates selective oligomerization. *J Biol Chem* **275**, 17173-9 (2000).
31. Babb, R. & Bowen, B.R. SDP1 is a peroxisome-proliferator-activated receptor gamma 2 co-activator that binds through its SCAN domain. *Biochem J* **370**, 719-27 (2003).
32. Inoue, M. et al. LDOC1, a novel MZF-1-interacting protein, induces apoptosis. *FEBS Lett* **579**, 604-8 (2005).
33. Baksh, S. et al. The tumor suppressor RASSF1A and MAP-1 link death receptor signaling to Bax conformational change and cell death. *Mol Cell* **18**, 637-50 (2005).
34. Tan, K.O. et al. MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. *J Biol Chem* **276**, 2802-7 (2001).
35. Okabe, H. et al. Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. *Cancer Res* **63**, 3043-8 (2003).

36. Lux, A. et al. Human retroviral gag- and gag-pol-like proteins interact with the transforming growth factor-beta receptor activin receptor-like kinase 1. *J Biol Chem* **280**, 8482-93 (2005).
37. Chih, D.Y. et al. Protein partners of C/EBPepsilon. *Exp Hematol* **32**, 1173-81 (2004).
38. Fujimoto, T., Tanaka, H., Kumamaru, E., Okamura, K. & Miki, N. Arc interacts with microtubules/microtubule-associated protein 2 and attenuates microtubule-associated protein 2 immunoreactivity in the dendrites. *J Neurosci Res* **76**, 51-63 (2004).
39. Lyford, G.L. et al. Arc, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites. *Neuron* **14**, 433-45 (1995).
40. Donai, H. et al. Interaction of Arc with CaM kinase II and stimulation of neurite extension by Arc in neuroblastoma cells expressing CaM kinase II. *Neurosci Res* **47**, 399-408 (2003).