**Supplemental Material**

**Figure S1. PTB- and IRS-containing proteins in eukaryotes.**

**A.** The human genome encodes 57 PTB-containing proteins (table S2), divided into two classes according to their domain sequence profile, the DAB/SHC-like PTB domain (43 proteins) and the IRS-like PTB (IRS) domain (14 proteins). Color code: gray, gene/protein present; white, gene/protein absent; black, not all diagnostic protein domains found or only a partial protein prediction is present in the database. Genes encoding proteins with PTB or IRS domains underwent a substantial expansion across the 31 organisms examined (table S1), but are absent in plants (e.g., the *Arabidopsis thaliana* control), as already reported (Manning, et al. 2008; Kaneko, et al. 2012). In *Dictyostelium discoideum*, there are two genes, Talin1- and Talin2-like genes, displaying an IRS-like PTB domain (Tsujioka, et al. 2008). The full set of human homologs appeared starting from cartilaginous fishes (Elephant shark), while in the genomes of jawless fishes (Sea Lamprey) many of the paralogs are not present, presumably due to the loss of one paralog, as has been reported for other gene families in this organism (Canestro, et al. 2013; Smith, et al. 2013). The choanoflagellate, *Monosiga brevicollis* and *Salpingoeca rosetta* possess a high number of PTB-containing proteins, but only a few of them are homologues to metazoan PTB-containing proteins (table S3), as already reported (Manning, et al. 2008). (i) The *Taeniopygia guttata* SH2D5 predicted protein (XP\_012425794) is a fusion gene between SH2D5 (PTB domain) and KIF17 (Kinesin motor Domain) and this is possibly due to an incomplete assembly of the genome. (ii) In Sea Lamprey (*Petromyzon marinus),* there are three NUMB paralogs, more similar to the mammalian NUMB than to NUMBL, which do not cluster with the vertebrate NUMB proteins. All three proteins are predicted from the genome with no evidence of expression (ESTs or mRNAs): ENSPMAP00000002145 lacks the starting methionine and part of the PTB domain (fig. 1 and fig. S6). (iii) The *Oikopleura dioica* genome encodes for two *NUMB* paralogs and their mRNAs are expressed. Both of them are more similar to the human NUMB than to human NUMBL (fig. S5). **B.** Number of duplicated or novel proteins. Novel proteins are those that do not resemble any of the human PTB or IRS-like PTB domain-containing proteins (See also table S3).

**Figure S2. Multiple alignment of NUMB and NUMBL proteins.**

NUMB and NUMBL proteins were aligned with ClustalX and the alignment was manually edited and visualized with Jalview. As indicated, the PTB domain is boxed in red, the NumbF domain in green and the proline-rich region (PRRi, coded by *NUMB* Ex9) in blue (see also fig. S3). Residues are colored according to default scheme in ClustalX. The actual NUMB protein of *Pleurobrachia bachei* was predicted from the *P. bachei* DRAFTgenome assembly (GCA\_000695325.1), using GeneWise and the *Mnemiopsis leidyi* NUMB protein as a template; the resulting predicted protein is longer, contains the starting methionine and is more accurate than the one present in the protein database (sb|3474499|).

**Figure S3. Ex3 and Ex9 of the *NUMB* gene are not present in *NUMBL*.**

**A**. Multiple alignment of vertebrate NUMB and NUMBL proteins, the latter highlighted in red. The PTB domain is boxed in red, the NumbF domain in green and the PPR1 (coded by *NUMB* Ex9) in blue. Exon boundaries and exon phases of human *NUMB* and *NUMBL* are indicated with arrowheads above or below the alignment, respectively. **B**. Exon structure and nucleotide length of vertebrate *NUMB* and *NUMBL* genes (N.A., not available; N.P., not present). Exons coding for the PTB domain are shaded in gray; \* = Stop codon.

**Figure S4. The NumbF domain is evolutionarily conserved in the NUMB protein family.**

Multiple alignment of NumbF domains of NUMB homologs. The predicted secondary structure of the human NUMB NumbF domain is depicted above the alignment. The NumbF domain is a region of unknown function, present only in the NUMB family of proteins (both NUMB and NUMBL), which has been shown to be involved in the binding to RAB7A/B and ERBB2 (Hirai, et al. 2017). The NumbF domain is well conserved in almost all NUMB and NUMBL proteins.

**Figure S5. *Oikopleura dioica* possesses two NUMB paralogs.**

**A-B**. The two *Oikopleura dioica* *NUMB* genes have different structures (tracks highlighted in yellow in the snapshot of the *Oikopleura dioica* Genome Browser <http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/>), and encode for two different proteins (24% identity) with more similarity to the human NUMB than to the human NUMBL. The Oikopleura ESTs, mapped on the two loci (tracks in red in the pictures), support the existence and the expression of the two mRNAs and proteins. **A**. NUMB GSOIDT00015583001 gene locus. **B**. NUMB-Like GSOIDT00008916001 gene locus. **C**. Alignment of the Oikopleura NUMB proteins with the human NUMB and NUMBL proteins; the PTB domain is boxed in red and the NumbF domain in green. **D**. Minimum Evolution Phylogenetic tree from the multiple alignment in C.

**Figure S6. Lamprey genomes encode for three NUMB paralogs.**

The genomic loci of the three *NUMB* genes encoded by the *Petromyzon marinus* (Sea Lamprey) genome are on different and distinct genomic contigs and are not products of an incorrect assembly. **A**. Visualization plots generated using mVISTA (Frazer, et al. 2004) to highlight conserved regions among the three Sea Lamprey *NUMB* gene loci. The plots show conserved sequences between *NUMB* Locus (GL478828:2703-29287) and NUMBL locus (GL476363:5139-17413) (Alignment 1) and *NUMB* Locus (GL478828:2703-29287) and *NUMB-Like* Locus (GL483700:1-8466) (Alignment 2) based on the multiple three-genome alignment using Multi-LAGAN (Brudno, et al. 2003). The level of conservation (vertical axis) is displayed in the coordinates of the sequence (horizontal axis). Conserved regions above the level of 70%/100 bp are highlighted under the curve. Predicted exons of the *Petromyzon marinus* *NUMB* gene (used as the reference in the alignments) are displayed above the graph. Note that only sequences corresponding to exons are similar, whereas intronic sequences are not. **B**.Multiple sequence alignment of the three NUMB proteins of *Petromyzon marinus* (Sea Lamprey Accession numbers: NUMB, ENSPMAP00000001971; NUMBL, ENSPMAP00000006450; NUMB-Like, ENSPMAP00000002145) and of *Lethenteron camtschaticum* (Japanese Lamprey Accession numbers: NUMB, JL4438; NUMBL, JL5974; NUMB-Like, JL2165) together with the human NUMB and NUMBL proteins; the PTB domain is boxed in red and the NumbF domain in green. Note that only the NUMB, and not the NUMBL or NUMB-Like proteins contain the long isoform of the PTB domain (PTB-long). **C**. Minimum Evolution Phylogenetic tree from the multiple alignment in B. The percentage of replicate trees in which the associated protein clustered together in the bootstrap test are shown next to the branches.

**Table S1. Complete list of organisms analyzed in this study.**

|  |  |
| --- | --- |
| ***Species name*** | **Common Name** |
| *Homo sapiens* | Human |
| *Gallus gallus* | Chicken |
| *Taeniopygia guttata* | Zebra finch |
| *Anolis carolinensis* | Lizard |
| *Xenopus tropicalis* | Western clawed frog |
| *Takifugu rubripes* | Japanese pufferfish (or Fugu) |
| *Danio rerio* | Zebrafish |
| *Callorhinchus milii* | Elephant Shark |
| *Petromyzon marinus* | Sea Lamprey |
| *Ciona intestinalis* | Sea Squirt |
| *Oikopleura dioica* | Oikopleura dioica |
| *Branchiostoma floridae* | Florida lancelet |
| *Strongylocentrotus purpuratus* | Sea Urchin |
| *Saccoglossus kowalevskii* | Acorn worm |
| *Capitella telata* | Polychaete worm |
| *Caenorhabditis elegans* | Caenorhabditis elegans |
| *Lottia gigantea* | Owl Limpet |
| *Drosophila melanogaster* | Fruit Fly |
| *Hydra vulgaris* | Hydra |
| *Nematostella vectensis* | Sea Anemone |
| *Trichoplax adhaerens* | Trichoplax adhaerens |
| *Amphimedon queenslandica* | Sponge |
| *Oscarella carmela* | Slime sponge |
| *Xestospongia testudinaria* | Barrel sponge |
| *Stylissa carteri* | Elephant ear sponge |
| *Mnemiopsis leidyi* | Sea walnut |
| *Pleurobrachia bachei* | Sea gooseberry |
| *Monosiga brevicollis* | Monosiga brevicollis |
| *Salpingoeca rosetta* | Salpingoeca rosetta |
| *Dictyostelium discoideum AX4* | Slime mold |
| *Arabidopsis thaliana* | Mouse-ear cress |

**Table S2. List of human PTB and IRS-like PTB domain-containing proteins.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Domain Organization** | **RefSeq Acc. numbers** | **Description** |
| **ANKS1A** | ANK-SAM-SAM-PTB | NP\_056060.2 | Ankyrin repeat and SAM domain-containing protein 1A |
| **ANKS1B** | ANK-SAM-SAM-PTB | NP\_690001.3 | Ankyrin repeat and sterile alpha motif domain-containing protein 1B |
| **APBA1** | PTB-PDZ-PDZ | NP\_001154.2 | Amyloid beta A4 precursor protein-binding family A member 1 |
| **APBA2** | PTB-PDZ-PDZ | NP\_005494.2 | Amyloid beta A4 precursor protein-binding family A member 2 |
| **APBA3** | PTB-PDZ-PDZ | NP\_004877.1 | Amyloid beta A4 precursor protein-binding family A member 3 |
| **APBB1** | WW-PTB-PTB | NP\_001155.1 | Amyloid beta A4 precursor protein-binding family B member 1 |
| **APBB2** | WW-PTB-PTB | NP\_004298.1 | Amyloid beta A4 precursor protein-binding family B member 2 |
| **APBB3** | WW-PTB-PTB | NP\_573420.2 | Amyloid beta A4 precursor protein-binding family B member 3 |
| **APPL1** | BAR-PH-PTB | NP\_036228.1 | Adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 1 |
| **APPL2** | BAR-PH-PTB | NP\_060641.2 | Adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 2 |
| **CCM2** | PTB-CCM2 | NP\_001025006.1 | CCM2 scaffolding protein |
| **CCM2L** | PTB | NP\_542192.2 | CCM2 Like scaffolding protein |
| **DAB1** | PTB | NP\_066566.3 | Disabled homolog 1 |
| **DAB2** | PTB | NP\_001334.2 | Disabled homolog 2 |
| **EPS8** | PTB-SH3 | NP\_004438.3 | Epidermal growth factor receptor kinase substrate 8 |
| **EPS8L1** | PTB-SH3 | NP\_573441.2 | Epidermal growth factor receptor kinase substrate 8-Like protein 1 |
| **EPS8L2** | PTB-SH3 | NP\_073609.2 | Epidermal growth factor receptor kinase substrate 8-Like protein 2 |
| **EPS8L3** | PTB-SH3 | NP\_620641.1 | Epidermal growth factor receptor kinase substrate 8-Like protein 3 |
| **FAM43A** | PTB | NP\_710157.2 | Protein FAM43A |
| **FAM43B** | PTB | NP\_997217.1 | Protein FAM43B |
| **GULP1** | PTB | NP\_057399.1 | PTB domain-containing engulfment adapter protein 1 |
| **ITGB1BP1** | PTB | NP\_004754.1 | Integrin beta-1-binding protein 1 |
| **LDLRAP1** | PTB | NP\_056442.2 | Low density lipoprotein receptor adapter protein 1 |
| **MAPK8IP1** | SH3-PTB | NP\_005447.1 | Mitogen-activated protein kinase 8 interacting protein 1 |
| **MAPK8IP2** | SH3-PTB | NP\_036456.1 | Mitogen-activated protein kinase 8 interacting protein 2 |
| **NOS1AP** | PTB | NP\_055512.1 | Nitric oxide synthase 1 adaptor protein |
| **NUMB** | PTB-NumbF | NP\_001005743.1 | Protein NUMB homolog |
| **NUMBL** | PTB-NumbF | NP\_004747.1 | NUMB Like protein |
| **PID1** | PTB | NP\_060403.3 | Phosphotyrosine interaction domain containing 1 |
| **RABGAP1** | PTB-TBC | NP\_036329.3 | Rab gtpase-activating protein 1 |
| **RABGAP1L** | PTB-TBC | NP\_055672.3 | Rab gtpase-activating protein 1-Like |
| **RGS12** | PDZ-PTB-RGS-RBD-RBD | NP\_937872.1 | Regulator of G-protein signaling 12 |
| **SH2D5** | PTB-SH2 | NP\_001096631.1 | SH2 domain-containing protein 5 |
| **SHC1** | PTB-SH2 | NP\_892113.4 | SHC-transforming protein 1 |
| **SHC2** | PTB-SH2 | NP\_036567.2 | SHC-transforming protein 2 |
| **SHC3** | PTB-SH2 | NP\_058544.3 | SHC-transforming protein 3 |
| **SHC4** | PTB-SH2 | NP\_976224.3 | SHC-transforming protein 4 |
| **TBC1D1** | PTB-PTB-TBC | NP\_055988.2 | TBC1 domain family member 1 |
| **TBC1D4** | PTB-PTB-TBC | NP\_055647.2 | TBC1 domain family member 4 |
| **TNS1** | PTEN\_C2-SH2-PTB | NP\_072174.3 | Tensin-1 |
| **TNS2** | C1-PTEN\_C2-SH2-PTB | NP\_056134.2 | Tensin-2 |
| **TNS3** | (C1)-PTEN\_C2-SH2-PTB | NP\_073585.8 | Tensin-3 |
| **TNS4** | SH2-PTB | NP\_116254.4 | Tensin-4 |
| **DOK1** | PH-IRS | NP\_001372.1 | Docking protein 1 |
| **DOK2** | PH-IRS | NP\_003965.2 | Docking protein 2 |
| **DOK3** | PH-IRS | NP\_079148.2 | Docking protein 3 |
| **DOK4** | PH-IRS | NP\_060580.2 | Docking protein 4 |
| **DOK5** | PH-IRS | NP\_060901.2 | Docking protein 5 |
| **DOK6** | PH-IRS | NP\_689934.2 | Docking protein 6 |
| **DOK7** | PH-IRS | NP\_775931.3 | Docking protein 7 |
| **FRS2** | IRS | NP\_006645.3 | Fibroblast growth factor receptor substrate 2 |
| **FRS3** | IRS | NP\_006644.1 | Fibroblast growth factor receptor substrate 3 |
| **IRS1** | PH-IRS | NP\_005535.1 | Insulin receptor substrate 1 |
| **IRS2** | PH-IRS | NP\_003740.2 | Insulin receptor substrate 2 |
| **IRS4** | PH-IRS | NP\_003595.1 | Insulin receptor substrate 4 |
| **TLN1** | B41-IRS-ILWEQ | NP\_006280.3 | Talin 1 |
| **TLN2** | B41-IRS-ILWEQ | NP\_055874.2 | Talin 2 |

**Table S3.** List of PTB- and IRS domain containing proteins found in the analyzed proteomes. Each separate sheet contains the detailed output of the HMMER (version 3.1b1) for each organism, together with accession numbers, protein domain composition and closest homologs identified by reverse BLAST search for each identified protein.

**References**

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721-731.

Canestro C, Albalat R, Irimia M, Garcia-Fernandez J. 2013. Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Semin Cell Dev Biol 24:83-94.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273-279.

Hirai M, Arita Y, McGlade CJ, Lee KF, Chen J, Evans SM. 2017. Adaptor proteins NUMB and NUMBL promote cell cycle withdrawal by targeting ERBB2 for degradation. J Clin Invest 127:569-582.

Kaneko T, Joshi R, Feller SM, Li SS. 2012. Phosphotyrosine recognition domains: the typical, the atypical and the versatile. Cell Commun Signal 10:32.

Manning G, Young SL, Miller WT, Zhai Y. 2008. The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. Proc Natl Acad Sci U S A 105:9674-9679.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, et al. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet 45:415-421, 421e411-412.

Tsujioka M, Yoshida K, Nagasaki A, Yonemura S, Muller-Taubenberger A, Uyeda TQ. 2008. Overlapping functions of the two talin homologues in Dictyostelium. Eukaryot Cell 7:906-916.