Hippo Pathway Phylogenetics Predicts
Monoubiquitylation of Salvador and Merlin/Nf2

Robert G. Wisotzkey1*, Charlotte E. Konikoff2*, Stuart J. Newfeld3*

1 NextBio, Santa Clara, California, United States of America, 2 Department of Biology, University of Washington, Seattle, Washington, United States of America, 3 School of Life Sciences, Arizona State University, Tempe, Arizona, United States of America

Abstract
Recently we employed phylogenetics to predict that the cellular interpretation of TGF-β signals is modulated by monoubiquitylation cycles affecting the Smad4 signal transducer/tumor suppressor. This prediction was subsequently validated by experiments in flies, frogs and mammalian cells. Here we apply a phylogenetic approach to the Hippo pathway and predict that two of its signal transducers, Salvador and Merlin/Nf2 (also a tumor suppressor) are regulated by monoubiquitylation. This regulatory mechanism does not lead to protein degradation but instead serves as a highly efficient “off/on” switch when the protein is subsequently deubiquitylated. Overall, our study shows that the creative application of phylogenetics can predict new roles for pathway components and new mechanisms for regulating intercellular signaling pathways.

Introduction
Phylogenetic analyses of amino acid conservation in developmental signaling pathways have traditionally focused on pathway origins [1] or evolutionary relationships within one/two families that encode pathway components [2]. Recently we stepped outside the box and applied phylogenetics to questions of pathway regulation. Our analyses of the TGF-β pathway predicted regulation by novel phosphorylation events and monoubiquitylation affecting Smad signal transducers [3,4]. Monoubiquitylation does not lead to protein degradation but instead serves as a highly efficient “off/on” switch when the protein is subsequently deubiquitylated [5]. Both of these predictions were validated by experiments in flies, frogs and mammalian cells demonstrating that the predicted regulatory mechanisms are fundamental features of TGF-β signaling [6,7,8,9].

Here we report a phylogenetic analysis of nine families that participate in the Hippo pathway and compare the results to the TGF-β (five families) and Wnt (ten families) pathways [10,11]. Our goal is to predict new regulatory mechanisms. Hippo pathway activation in flies (Figure 1A left) begins when a signal from a neighboring cell is communicated through two atypical transmembrane cadherins - the ligand Dachsous and its receptor Fat. Fat then initiates two cascades that converge synergistically downstream. Fat blocks the activity of the myosin-like protein Dachs by preventing it from accumulating at the membrane. Inhibition of Dachs leads to increased stability of the Warts serine-threonine kinase. Fat also increases the membrane association of the Ferm domain protein Expanded that leads to an increase in the activity of Warts. Expanded accomplishes this via a complex containing the Ferm domain protein Merlin and the WW domain protein Kibra. This complex then recruits the Hippo kinase complex (Hippo, Salvador, Warts and Mats) to the membrane where Hippo and Salvador undergo phosphorylation via an unknown mechanism.

The Hippo serine-threonine kinase, assisted by its co-factor Salvador, then phosphorylates Warts and its co-factor Mats. Salvador and Mats are WW/Sarah and Phoecin domain adapter proteins, respectively. Phosphorylated Warts then phosphorylates Yorkie, a WW domain transcription co-factor. Phosphorylated Yorkie is recognized and sequestered in the cytoplasm by 14-3-3 proteins. In the absence of phosphorylation Yorkie continually translocates to the nucleus, binds the transcription factor Scalloped and influences gene expression that leads to increases in cell proliferation and decreased apoptosis [12]. In vertebrates (Figure 1A right), Hippo signaling is similar and elicits similar outcomes, but not all homologs of fly Hippo pathway genes have been implicated in the vertebrate Hippo pathway [13].

Previously, a phylogenetic study of the Yorkie/Yap/Taz family and its binding partners in the Scalloped/Tead family revealed evidence of co-evolution [2]. Another study focused on pathway origins showed that the Hippo/Mst, Warts/Lats, Yorkie/Yap/Taz and Scalloped/Tead families are present in a unicellular amoeboid species and that the amoeboid proteins can promote tissue growth in Drosophila [1]. The phylogenetics data we present have a larger scope then either of these studies: we analyze nine Hippo pathway families and compare the results to data for TGF-β and Wnt pathway families.

Three lines of phylogenetic data converge to support predictions that two non-enzymatic Hippo pathway signal transducers,
Results and Discussion

Experimental Design for Phylogenetic Predictions of Lysine Monoubiquitylation

Phylogenetic analyses of amino acid alignments have successfully improved our understanding of TGF-β pathway regulation. The underlying logic of the phylogenetic approach is to exploit evolutionary conservation as a guide to amino acids involved in the post-translational regulation of protein activity. As both monou- and polyubiquitylation occur solely on lysine (with very rare exceptions) the identification of potential monoubiquitylation sites requires consulting available biochemical data to eliminate lysines conserved for structural reasons. Monoubiquitylation predictions for a specific lysine are then based on: 1) universal conservation in a broad range of species, 2) occurrence within an highly conserved context and 3) that is not essential for protein structure [4].

In this study we also required that the conserved context include an upstream hydrophobic residue (at positions -1 or -2) as we noted for the two monoubiquitylated lysines (K319 and K507) in human Smad4 [8,14]. The possibility that an upstream hydrophobic residue is part of a monoubiquitylation “signature” is supported by several computational analyses. These showed that the context is a significant factor in ubiquitylation [15], that hydrophathy is the most important feature of the context for ubiquitylation [16] and that polyubiquitylation is associated with a statistically significant depletion of the hydrophobic residue leucine at positions -1 and -2 [17]. Thus the presence of a hydrophobic amino acid in these positions may allow the ubiquitin ligase to discriminate between poly- and monoubiquitylation.

The nine highly divergent species employed in this study and their Linnaean classifications are shown in Figure 1B. All are coelomates with three embryonic germ layers and a digestive tract with two openings: eight are deuterostomes (blastopore becomes the anus) and one is a protostome (blastopore becomes the mouth). Deuterostomes and protostomes diverged 1.1 billion years ago. Within the deuterostomes, echinoderms (sea urchins) and hemichordates (acorn worms) diverged 990 million years ago (mya) from chordates. Within the chordates, cephalochordates (lancelets) and urochordates (ascidians) diverged 750 and 720 mya from vertebrates. Within the vertebrates, zebrafish diverged 750 and 720 mya from sheep. Within mammals, mice diverged from humans 93 mya [18].

Salvador and Merlin/Nf2 Lysine Conservation Resembles Smad4 Lysine Conservation

In the Hippo family, human Mst1 contains thirty-six lysines (Table 1). Eighteen (50%) of these are universally conserved but biochemical data suggests that fifteen are structural [19]. We examined these fifteen lysine residues according to the criteria noted above (universal conservation, conserved context, upstream hydrophobic residue and not structural) and estimated a qualitative likelihood of monoubiquitylation. Based on a hydrophobic leucine at position -1 in all species and its location upstream of the kinase domain we predict that Mst1 K15 is a modest candidate for monoubiquitylation.

Human Salvador1 contains eighteen lysines (Table 1) of which only one (5.6%) is universally conserved. Salvador K345 occurs within the Sarah domain that binds Hippo/Mst1 but biochemical data suggests it is not structural [20]. K345 monoubiquitylation could block Salvador/Hippo complex formation just as Smad4 monoubiquitylation at K319 blocks Smad4/Smad2 binding [8]. In addition, Salvador K345 has a hydrophobic methionine at -1 in vertebrates and a leucine at -2 in all other species. The similarity
between Smad4 K519 and Salvador K345 led us to predict that K345 is a strong candidate for monoubiquitylation. Differences in lysine conservation between Salvador and the polyubiquitylated Armadillo/β-catenin family are consistent with the prediction of Salvador K345 monoubiquitylation. Human β- catenin contains twenty-six lysines of which three (11.5%) are universally conserved. These lysines are structural components of the interaction surface that binds the transcription factor TCF [21]. Further, in contrast to the monoubiquitylated K519 of Smad4 that is conserved in C. elegans as sma-4 K538, the polyubiquitylated K19 in human β-catenin is not conserved in C. elegans wvm-1 [10]. Since polyubiquitylation need not occur within an interaction domain, the target lysine is not always conserved.

In the Warts family, human Lats1 contains sixty-four lysines (Table 1) of which fourteen (21.9%) are universally conserved. Based on the structure of the related Akt serine-threonine kinase, ten of these appear structural [22]. Lats1 K652 occurs in the Yorkie/Yap/Taz interaction domain, has a hydrophobic residue at -1 in all species and is predicted to be ubiquitylated by the UbiPred algorithm [17]. We consider K652 a strong monoubiquitylation candidate. In the Mats family, human Mob1α contains seventeen lysines (Table 1) of which eight (47.1%) are universally conserved and three are structural [23]. Mob1α K104 is located upstream of the dimerization interface, has a hydrophobic residue at -1 in all species and is predicted by UbiPred. We consider K104a strong monoubiquitylation candidate.

In the Yorkie family, human Yap1 has fourteen lysines (Table 1) of which none (0%) are universally conserved. The absence of lysine conservation is unparalleled among the twenty-four families of the Hippo, TGF-β and Wnt pathways we have studied [4,10]. Two human Yap1 lysines (K90 and K204) are conserved within vertebrates and since K90 has a leucine at -1 and K204 is predicted by UbiPred, we consider these modest monoubiquitylation candidates. Consistent with the β-catenin data, K97 and K242 that were shown to be polyubiquitylated in mouse Yap1 [24,25] are not conserved outside mammals. In the Scalloped family, human Tead1 has twenty-eight lysines (Table 1) of which twelve (42.9%) are universally conserved and six are structural [26]. Tead1 K289 is located in the Yorkie/Yap/Taz interacting region, has a leucine at -2 and a valine at -1 in all species and is predicted by UbiPred. We consider Tead1 K289 a strong monoubiquitylation candidate.

In the Kibra family, human Wwc1 has seventy-four lysines (Table 1) of which only one (1.4%) is universally conserved. Wwc1 K43 has a leucine at -2 in vertebrates and is located within a WW binding domain. As no structural data on human Wwc1 is available, we consider K43 a modest candidate for monoubiquitylation in vertebrates.

In the Expanded family, human Frmd1 has twenty lysines (Table 1) of which seven (35%) are universally conserved and two are structural [27,28]. Frmd1 K116 has a hydrophobic residue at -2 in all species, is located within a Ferm interaction domain and is predicted by UbiPred. We consider K116 a strong candidate for monoubiquitylation.

In the Merlin family, human Nf2 has sixty lysines (Table 1) with seven (11.7%) universally conserved and three structural [27,28]. K322 has a conserved hydrophobic residue at -1, is located between lipid binding and ERM domains and is predicted by UbiPred. We consider K322 a strong candidate for monoubiquitylation. After making this prediction, we found evidence in the literature suggesting that K322 is monoubiquitylated [29]. In that paper, an N-terminal fragment containing residues 1–322 of human Nf2 was monoubiquitylated in cell culture by an endogenous ligase (as shown in that paper’s Figure 4E upper right panel lane 2). Upon replacing a serine or threonine with a phosphomimetic glutamic acid (T230D and/or S315D) this fragment became polyubiquitylated (as shown in that paper’s Figure 4E upper right panel lanes 3,4).

---

**Table 1.** Summary of absolutely conserved lysine residues and predictions for monoubiquitylation candidates in human Hippo pathway proteins.

<table>
<thead>
<tr>
<th>Protein Family</th>
<th>Conserved lysines</th>
<th>Total lysines</th>
<th>Conserved fraction</th>
<th>Monoubiquitylation candidate lysine*</th>
<th>Conserved contextb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hippo/Mst</td>
<td>18 Hs_Mst1</td>
<td>36</td>
<td>50%</td>
<td>K15 just upstream of the kinase domain is a modest candidate</td>
<td>L-K-K-L-S/D in all species</td>
</tr>
<tr>
<td>Salvador</td>
<td>1 Hs_Sav1</td>
<td>18</td>
<td>5.6%</td>
<td>K345 in the Hippo/Mst1 binding domain is a strong candidate</td>
<td>M-K-E-L-E in vertebrates</td>
</tr>
<tr>
<td>Warts/Lats</td>
<td>14 Hs_Lats1</td>
<td>64</td>
<td>21.9%</td>
<td>K652 in the Yap/Taz binding domain is a strong candidate</td>
<td>I/L/M/K-S/T in all species</td>
</tr>
<tr>
<td>Mats/Mob</td>
<td>8 Hs_Mob1a</td>
<td>17</td>
<td>47.1%</td>
<td>K104 just upstream of the dimerization domain is a strong candidate</td>
<td>I/V-K-K-P-I in all species</td>
</tr>
<tr>
<td>Yorkie/Yap/Taz</td>
<td>0 Hs_Yap1</td>
<td>14</td>
<td>0%</td>
<td>K90 in the Scalloped/Tead binding domain is a modest candidate</td>
<td>L-R-K-L in vertebrates</td>
</tr>
<tr>
<td>Scalloped/Tead</td>
<td>12 Hs_Tead1</td>
<td>28</td>
<td>42.9%</td>
<td>K289 in the Yap/Taz binding domain is a strong candidate</td>
<td>L-V-K-F-W-A in all species</td>
</tr>
<tr>
<td>Kibra/Wwc</td>
<td>1 Hs_Wwc1</td>
<td>74</td>
<td>1.4%</td>
<td>K43 between two Wwc binding domains is a modest candidate</td>
<td>L-T-K-P in vertebrates</td>
</tr>
<tr>
<td>Expanded/Frmd</td>
<td>7 Hs_Frmd1</td>
<td>20</td>
<td>35%</td>
<td>K116 in the first Ferm binding domain is a strong candidate</td>
<td>I/L-S/Y-E-Y in all species</td>
</tr>
<tr>
<td>Merlin/Nf2</td>
<td>7 Hs_Nf2</td>
<td>60</td>
<td>11.7%</td>
<td>K322 between lipid binding and ERM domains is a strong candidate</td>
<td>M-K-A-Q-A-R in all species</td>
</tr>
</tbody>
</table>

*If only one conserved lysine is present, then we consider it a monoubiquitylation candidate. If multiple conserved lysines are present, then we identify a candidate for monoubiquitylation (underlined). Qualitative assessment of the likelihood that the lysine is monoubiquitylated is indicated (strong or modest).

**Absolutely conserved amino acids surrounding the candidate lysine (underlined) are shown. If amino acids are separated by a slash then that position is variable and the amino acids present are shown. A hydrophobic amino acid (I, L, V or M) in the -1 or -2 position is shown in **bold.**

doi:10.1371/journal.pone.0051599.t001
targeted lysines were identified. That paper’s results strongly support our prediction because monoubiquitylation, as noted above, favors conserved lysines and K322 is the most upstream of the seven conserved lysines in Merlin/Nf2 (there is no other conserved lysine within 1–322) and polyubiquitylation favors unconserved lysines (there are twenty-five unconserved lysines within 1–322). Note that Merlin/Nf2 has roles in cell fate specification that are not connected to canonical Hippo signaling [30] and this prediction does not allow us to specify which of Merlin/Nf2’s functions may be regulated by monoubiquitylation.

In summary, the lysine conservation data most strongly predicts the monoubiquitylation of Salvador and Merlin/Nf2. For Salvador, this is based on the presence of only one conserved lysine that is not required structurally and that is present in a conserved context containing a hydrophobic residue at -1 or -2 in all species. For Merlin/Nf2, where suggestive experimental data exists in the literature, this is based on the presence of a conserved lysine that is not required structurally and that is present in a conserved context containing a hydrophobic residue at -1 in all species.

Salvador and Merlin/Nf2 Copy Number Reduction in Mammals Mimics Smad4 Copy Number Reduction

We then sought to identify additional parallels between Hippo pathway families and monoubiquitylated members of the Co-Smad subfamily (Smad4 in vertebrates and Medea in flies). A survey showed that vertebrate species in all Hippo families contain a variable number of family members.

In the Hippo/Mst family (Figure 2A) there are two members in each amniote (mammals and chicken). This is typical due to a vertebrate-specific whole-genome duplication that occurred in the Ordovician era [31]. However, in fish only Mst2 is present which is surprising since teleost fish experienced another whole-genome duplication after their divergence from other vertebrates in the Silurian era [32]. Thus in fish there are normally four copies of each gene/protein, but in the Hippo/Mst family fish have lost three copies. In the Salvador family there is only one member in each vertebrate (Figure 2B). This indicates loss of one copy in each amniote and three from fish. In the Warts/Lats family (Figure 3A) amniotes have the expected number but fish have lost two copies. In the Mats/Mob family (Figure 3B) mammals have the expected number but chicken and fish each have lost one copy. In the Yorkie/Yap/Taz family (Figure 4A) mammals have the expected number, but chicken has lost one and fish have lost two copies.

The Scalloped/Tead family (Figure 4B) has a complex history. Each mammal has four members. Tead2 is unique to mammals as chicken and fish each have only three copies. Chicken has Tead1, 3 and 4 while fish have two Tead1s and Tead3. The mammal specific Tead2 cluster is statistically linked to the cluster from the Mst/Mst3 cluster and both clusters form a group with Tead1. Both Tead3 and Tead1 are present in all vertebrates. The most parsimonious explanation for this distribution of Tead family members is that the ancestral vertebrate had four Tead genes as seen in mammals. Tead2 was lost in chicken and fish lost five Tead genes, a total of six events. Explaining the current distribution by ascribing two Tead genes to the ancestral vertebrate or by assigning Tead2 to a mammal-specific duplication requires a larger number of events. In the Kibra/WWc1 family (Figure 5A) amniotes have the expected number but fish have lost three copies. In the Expanded/Frdm family (Figure 5B) amniotes may have the expected number (see legend for a discussion of M. musculus Frmd1) but fish have lost two copies. In the Merlin/Nf2 family (Figure 5C) amniotes have each lost one copy while fish have lost two.

Figure 2. Hippo/Mst and Salvador Maximum Likelihood trees. Trees rooted with D. melanogaster are shown. Branches considered statistically weak (aLRT values <0.70) are colored in blue. Branch lengths are drawn to scale and correspond to the average number of amino acid changes per site as indicated by the scale bar. A) Hippo/Mst vertebrate topology matches the species tree. Among invertebrates the hemichordate strongly clusters with vertebrates while the echinoderm is an outlier. The overall Hippo/Mst tree matches the species tree. The Bayesian tree (Figure S1A) shows one difference - the cephalochordate moves out of the strong cluster with vertebrates to become an outlier like the echinoderm. B) Salvador vertebrate topology matches the species tree. Among invertebrates the urochordate clusters tightly with vertebrates while the echinoderm and cephalochordate are outliers. The overall Salvador tree matches the species tree. The Bayesian tree (Figure S1B) shows one difference - the cephalochordate clusters outside the urochordate and vertebrate cluster with confidence. doi:10.1371/journal.pone.0051599.g002

We then compared fly, mammal and fish copy number across the three pathways. As noted above, the expectation is two mammal and four fish genes/proteins for every one in flies. In mammalian TGF-β and Wnt, but not Hippo families, gene gain beyond this expectation is visible. For example: Smad1, 5, 8 are homologs of fly Mad [4] and Dvl1, 2, 3 are homologs of fly Dishevelled [33]. Alternatively, all pathways display gene loss in mammals. For Hippo, mammalian gene loss is seen in two of nine families (Salvador and Merlin/Nf2). For TGF-β (five families), mammalian gene loss is visible in the Co-Smad subfamily (Medea/Smad4). For Wnt (ten families), mammalian gene loss is visible in the β-catenin family.

Evidence of gene loss in Hippo families is also seen in fish. Hippo families in fish have universally experienced gene loss (down to a single copy in three families - Hippo/Mst, Salvador and Kibra/Wwc). Alternatively, many TGF-β and Wnt families in fish contain four or five copies reflecting both rounds of genome
duplication. For example, Dishevelled is present in five copies and Smad2 is present in four copies [33,34].

There is only one protein family in each of the TGF-β and Wnt pathways that is reduced to one copy in mammals and both of these proteins have pleiotropic functions. Smad4 belongs to the Co-Smad subfamily because it forms transcriptional complexes with many partners (Smads1, 2, 3, 5 and 8) and thus acts as a signal transducer for multiple TGF-β pathways. β-catenin family members play roles in Wnt signaling and in adherens junctions where they modulate the turnover of membrane-bound cadherins [10]. One possible explanation for copy number reduction in these multi-functional families is that this prevents unintended crosstalk between their various roles. Evolutionary mechanisms, such as adaptive mutations that prevent crosstalk are believed to strongly influence pathway diversification [35].

The logic that copy number reduction serves to prevent pathway crosstalk for Smad4 seems apt for Salvador. The sole copy of Salvador in vertebrates is thought as a co-factor for both Hippo/Mst kinases [11] just as the sole copy of Smad4 forms complexes with multiple Smad partners. Extending the similarity between these proteins to their roles as passive co-factors also seems reasonable. However, recently the view of Smad4 as a passive partner has been modified based on two reports: 1) that Smad4 monoubiquitylation cycles serves as highly efficient "off/on" switches for TGF-β signaling [8,9] and 2) that the fly Co-Smad Medea can be shuttled by cell type specific protein-protein interactions between Smad partners as an "inter-pathway" switch [36]. The similarities between Smad4 and Salvador in copy number and lysine conservation suggest at Salvador may function as an "off/on" switch for Hippo signaling via monoubiquitylation or as an "inter-pathway" switch between Mst1 and Mst2.

The logic that copy number reduction serves to prevent pathway crosstalk for Smad4 also seems apt for Merlin/Nf2. The sole copy of Merlin/Nf2 is thought to function primarily via

Figure 3. Warts/Lats and Mats/Mob Maximum Likelihood trees. A) Warts/Lats vertebrate topology matches the species tree. Among invertebrates, these proteins should be sequentially joined to the vertebrate cluster and thus this part of the Warts/Lats tree deviates from the species tree. The Bayesian tree (Figure S1C) shows one difference - rather than the hemichordate and echinoderm clustering together they are sequentially clustered with the large urochordate, cephalochordate and vertebrate group, an arrangement that better matches the species tree. B) Mats/Mob vertebrate topology matches the species tree except for difficulty resolving human, mouse and chicken Mob1a. Among invertebrates a cluster of the echinoderm and urochordate sequences is attached to the vertebrate cluster with the hemichordate and cephalochordates as outliers. The association of urochordates with vertebrates matches the species tree but the inclusion of echinoderms without hemichordates and cephalochordates does not. The Bayesian tree (Figure S1D) shows one difference - the urochordate switches places with the hemichordate leading to an arrangement that better matches the species tree but with no confidence.

doi:10.1371/journal.pone.0051599.g003

Figure 4. Yorkie/Yap/Taz and Scalloped/Tead Maximum Likelihood trees. A) Yorkie/Yap/Taz (Taz is formally known as Wwtr1) vertebrate topology matches the species tree. Among invertebrates there is strong pairing of the echinoderm and hemichordate but weak connections of this pair to cephalochordates, urochordates and vertebrates in relationships that deviate from the species tree. The Bayesian tree (Figure S1E) shows one difference - the urochordate has switched places with the cephalochordate exaggerating the deviation from the species tree. B) Scalloped/Tead vertebrate topology matches the species tree. Among invertebrates, the tree clusters the hemichordate and echinoderm with vertebrate Tead1 without the urochordate and clusters the urochordate with vertebrate Tead3, both are deviations from the species tree. The Bayesian tree (Figure S1F) shows one difference - the echinoderm and hemichordate move away from vertebrate Tead1 as outliers, an arrangement that is a slightly better match to the species tree.

doi:10.1371/journal.pone.0051599.g004
between Smad4 and Merlin/Nf2 in copy number and lysine conservation suggest that Merlin/Nf2 may also be regulated via monoubiquitylation (as suggested by existing experimental data) or may play a role as an “inter-pathway” switch between Frmd1–Fermd6 or Wwc1–Wwc2 or Fermd and Wwc pathways.

**Salvador and Merlin/Nf2 Discordant Partner Substitution Rates are Similar to Smad4 Discordant Partner Substitution Rates**

Unlike copy number and lysine conservation data that have small variable ranges, amino acid substitution rates in animals can vary up to 67,000-fold. For example, histone H4 shows a rate of 0.001 substitutions per residue per billion years (all substitution rate data is stated using this standard metric and abbreviated as substitutions) while ATP-cone domain proteins show a rate of 66.6 substitutions [37,38]. The latter rate is considered as fast as possible while still maintaining recognizable family membership.

First we calculated overall rates (Table 2) for six TGF-β families (range 0.258–1.339; average 0.861) and eleven Wnt families (range 0.333–1.730; average 1.184). We include the TGF-β and Wnt ligand families here, but not in the copy number analysis above since there is no a priori reason to expect a bias in substitution rates for ligands versus signal transducers whereas there is a well documented bias towards increasing copy number in ligands [39]. Overall, Hippo families are evolving more slowly than the other pathways (range 0.151–1.379; average 0.599).

When examining interacting proteins with similar roles in the TGF-β and Wnt pathways we frequently saw comparable substitutions rates. Concordant rates suggest the proteins are under similar selective constraints. Examples of interacting proteins with concordant substitution rates are: 1) the TGF-β receptor complex components (Type I 0.830, Type II 1.153), 2) Wnt receptor complex components (Fz 1.024, LRP 1.505) and 3) Wnt transcription factor complexes (β-catenin 1.244, TCF 1.021). Alternatively interacting proteins in which each plays a distinct role frequently showed discordant rates (differing by more than 2-fold). An example in the TGF-β pathway is the Smad multimeric complex composed of Co-Smads regulated by monoubiquitylation (0.595) and receptor activated Smads regulated by phosphorylation (0.258). An example in the Wnt pathway is the multimeric complex composed of Shaggy/Gsk3 kinases (0.333) with the adapter proteins APC (1.221) and Axin (1.583).

In the Hippo pathway, we found examples of concordance such as: Warts/Lats kinase (0.298) and its co-factor Mats/Mob (0.151) as well as Kibra/Wwc (0.802) and Expanded/Frmd (0.997). Discordance in substitution rates between interacting proteins is evident for Hippo/Mst (0.156) and Salvador (.821; over five fold higher) as well as Merlin/Nf2 (0.498) and Expanded/Frmd (0.997).

These findings are not expected from Salvador’s perceived role as a passive co-factor for the Hippo kinase or from the perceived roles of Merlin/Nf2 and Expanded/Frmd as simple adapter proteins. In summary, comparisons with discordant pairs in the TGF-β and Wnt pathways suggest that Salvador and Merlin/Nf2 have undiscovered roles distinct from their partners (e.g., monoubiquitylation), within the Hippo pathway or beyond.

Overall, this study builds upon successful phylogenetics-based predictions for the regulation of Smad signal transducers, including Smad4, in the TGF-β pathway. Comparison of phylogenetic data for nine Hippo families to results from studies of the TGF-β and Wnt pathways led us to predict new roles for Salvador and Merlin/Nf2 and to suggest that these may involve pathway regulation via monoubiquitylation. If validated, and suggestive experimental evidence is already visible in the literature for Merlin/Nf2, these predictions will reinforce the view that the
creative exploitation of amino acid conservation via phylogenetics can illuminate new regulatory mechanisms affecting developmental pathways.

Materials and Methods

Sequences

All members of the Hippo/Mats, Salvador, Warts/Lats, Mats/Mob, Yorkie/Yap/Taz, Scalloped/Tead, Kibra/Wwc, Expanded/Frmd and Merlin/Nf2 families for nine fully sequenced organisms were retrieved and analyzed as described [10]. The longest isoform was utilized and partial sequences excluded. Sequences are named according to their genus and species. See Table S1 for accession numbers. Calculations of the overall amino acid substitution rate for each family were conducted in MEGA5 using the Poisson correction model [40].

Alignments and Trees

Alignments for each gene family were created in MAFFT. In one case an alignment was annotated manually (in the Expanded/Frmd family an excessive tail on Expanded was removed). See Table S2 for detailed alignment length information. Maximum Likelihood trees (main text) were generated from alignments as described [10]. Bayesian trees (Figure S1) were generated from sequences aligned with MUSCLE [41] and created in MrBayes 3.1.1 [42]. For Bayesian trees, the prior amino acid model was set to Blosum. Gamma plus invariant sites were taken into account, with the gamma shape parameter being composed of four discrete categories. According to ProtTest [43], the gamma plus invariant model was the best-fit model for protein evolution in this family. The number of generations was set to 100,000 with a sample frequency of 100 and burn-in frequency of 0.25. This was the smallest number of generations that produced meaningful trees. All other parameters were set to default values.

Supporting Information

Figure S1 Bayesian trees of Hippo pathway proteins. Bayesian trees are displayed in the same format as the Maximum Likelihood trees in the main text with explicit posterior probabilities shown at the nodes. Branches with posterior probabilities <0.90 are considered weak and colored in blue. Branch length denotes the number of amino acid changes per site. Organisms are abbreviated as in the Maximum Likelihood trees in the main text. A) Hippo/Mst Bayesian tree, B) Salvador Bayesian tree, C) Warts/Lats Bayesian tree, D) Mats/Mob Bayesian tree, E) Yorkie/Yap/Wwtr Bayesian tree, F) Scalloped/Tead Bayesian tree, G) Kibra/Wwc Bayesian tree, H) Expanded/Frmd Bayesian tree and I) Merlin/Nf2 Bayesian tree.

Table S1 Accession numbers.

Table S2 Alignment lengths. The total lengths of protein alignments used to create trees are shown below. Bayesian trees were made from MUSCLE alignments, while MAFFT alignments were used to create Maximum Likelihood trees. Expanded alignments were corrected by hand since the presence of numerous large D. melanogaster insertions and deletions resulted in many gaps, making it difficult to create trees. Original alignment lengths for Ex are shown in parentheses. Alignments are available upon request.

Acknowledgments

We thank Sirio Dupont, Sudhir Kumar and Stefano Piccolo for valuable discussions.

Author Contributions

Conceived and designed the experiments: SJN. Performed the experiments: CEK RGW SJN. Analyzed the data: CEK RGW SJN. Wrote the paper: SJN.

References


Fig. S1A. Hippo/Mst Bayesian tree
Fig. S1B. Salvador Bayesian tree

Fig. S1C. Warts/Lats Bayesian tree
Fig. S1D. Mats/Mob Bayesian tree

Fig. S1E. Yorkie/Yap/Wwtr Bayesian tree
Fig. S1F. Scalloped/Tead Bayesian tree

Fig. S1G. Kibra/Wwc Bayesian tree
Fig. S1H. Expanded/Frmd Bayesian tree

Fig. S1I. Merlin/Nf2 Bayesian tree
Supplemental information

Table S1. Accession numbers

Hippo/Mst family: Dm_Hpo, NP_611427.1; Hs_Mst1, NP_006273.1; Hs_Mst2, NP_006272.2; Mm_Mst1, NP_067395.1; Mm_Mst2, NP_062609.2; Gg_Mst1, NP_001026024.1; Gg_Mst2, NP_001026508.1; Dr_Mst2, NP_955966.1; Sk_Mst3l, XP_002736706.1 and Sp_Krs2l, XP_781787.2. A partial sequence for B. floridae was found (XP_002589064.1) but excluded. Reciprocal Blasting revealed D. rerio Mst2 was also the best hit when searching with Mst1. Additional D. rerio hits with Mst1 or Mst2 were more closely related to other serine-threonine kinases. No strong hits were found for C. intestinalis.

Salvador family: Dm_Sav, NP_788721.1; Hs_Sav1, NP_068590.1; Mm_Sav1, NP_071311.1; Gg_Sav1, XP_421467.2; Dr_Sav1, NP_001004560.1; Bf_Sav, XP_002610502.1; Ci_Sav1, XP_002126709.1 and Sp_Sav XP_001178982.1. No strong hits were found for S. kowalevskii.

Warts/Lats family: Dm_Wts, NP_733403.1; Hs_Lats1, NP_004681.1; Hs_Lats2, NP_055387.2; Mm_Lats1, NP_034820.1; Mm_Lats2, NP_056586.2; Gg_Lats1, XP_419666.2; Gg_Lats2, XP_417143.2; Dr_Lats1, NP_001018346.1; Dr_Lats2, NP_001121728.1; Bf_Wts, XP_002587173.1; Ci_Wts, XP_002124396.1; Sk_Wts, XP_002738496.1 and Sp_Wts, XP_795100.2.

Mats/Mob family: Dm_Mats, NP_651041.3; Hs_Mob1a, NP_775739.1; Hs_Mob1b, NP_060691.2; Mm_Mob1a, NP_081011.1; Mm_Mob1b, NP_663546.1; Gg_Mob1a, XP_420601.1; Dr_Mob1aa, NP_956494.1; Dr_Mob1ab, NP_999948.2; Dr_Mob1b, NP_956208.1; Bf_Mats, XP_002596178.1; Ci_Mats, XP_002122354.1; Sk_Mats, XP_002741421.1 and Sp_Mats, XP_788775.1.

Yorkie/Yap/Wwtr family: Dm_Yki, NP_001036568.2; Hs_Yap1, NP_001123617.1; Hs_Wwtr1, NP_056287.1; Mm_Yap1, NP_033560.1; Mm_Wwtr1, NP_598545.2; Gg_Yki, NP_990574.1; Dr_Yap1, NP_001132952.1; Dr_Wwtr1, NP_001032785.1; Bf_Yki, XP_002595229.1; Ci_Yki, XP_002130260.1; Sk_Yap1l, XP_002734168.1 and Sp_Yap1l,
Scalloped/Tead family: Dm_Sd, NP_511169.1; Hs_Tead1, NP_068780.2; Hs_Tead2, NP_003589.1; Hs_Tead3, NP_003205.2; Hs_Tead4, NP_003204.2; Mm_Tead1, NP_033372.1; Mm_Tead2, NP_035695.1; Mm_Tead3, NP_001091696.2; Mm_Tead4, NP_035697.3; Gg_Tead1, XP_420962.2; Gg_Tead3, XP_419267.1; Gg_Tead4, NP_001001339.1; Dr_Tead1, XP_001345193.2; Dr_Tead1a, NP_998012.1; Dr_Tead3, NP_001103194.1; Ci_Tead1, NP_001087209.1; Sk_Sd, NP_001161650.1 and Sp_Tead1, XP_785374.2. A partial sequence for *B. floridiae* was also found (XP_002595807.1), but was excluded.

Kibra/Ww family: Dm_Kibra, NP_001034055.1; Hs_Wwc1, NP_056053.1; Hs_Wwc2, NP_079225.5; Mm_Wwc1, NP_740749.1; Mm_Wwc2, NP_598552.2; Gg_Wwc1, XP_414499.2; Gg_Wwc2, XP_420516.2; Dr_Wwc1l, XP_689275.3; Bf_Kibra, XP_002598149.1; Ci_Kibra, XP_002129513.1 and Sp_Kibra, XP_780710.2. Dm Kibra is also called CG33967. Some *D. rerio* sequences with high similarity via reciprocal blasting were discontinued in GenBank and thus excluded. No strong hits were found for *S. kowalevskii*.

Expanded/Fermd family: Dm_Ex, NP_476840.2; Hs_Frmd1, NP_079195.3; Hs_Frmd6, NP_001035946.1; Mm_Frmd6, NP_082403.2; Gg_Frmd1, XP_419603.2; Gg_Frmd6, XP_426484.2; Dr_Frmd6, XP_002664244.1; Dr_Frmd6l, NP_001104657.1; Bf_Ex, XP_002590857.1 and Sp_Frmd6l, XP_797238.2. A sequence for *M. musculus* Frmd1 is present, but was excluded because it appeared to have a translation discrepancy (LOC544719). No strong hits were found in *S. kowalevskii*, though many other ezrin/radixin/moesin members are clearly present. Two potential tandem *C. intestinalis* sequences (XP_002125027.1 and XP_002125095.1) were excluded as we took a conservative approach and only analyzed sequences with significant e-value matches to the query sequence.

Merlin/Nf2 family: Dm_Mer, NP_523413.1; Hs_Nf2, NP_000259.1; Mm_Nf2, NP_035028.2; Gg_Nf2, NP_989828.2; Dr_Nf2a, NP_001122179.1; Dr_Nf2b, NP_998116.1; Bf_Nf2, XP_002591890.1; Ci_Mer, XP_002125990.1; Mm_Mer, BAF49216.1; Sk_Mer, NP_001164711.1 and Sp_Nf2, XP_781142.2.
Table S2. Alignment lengths. The total lengths of protein alignments used to create trees are shown below. Bayesian trees were made from MUSCLE alignments, while MAFFT alignments were used to create Maximum Likelihood trees. Expanded alignments were corrected by hand as many long D. melanogaster indels resulted in a large number of gaps, making it difficult to create trees. Original alignment lengths for Ex are shown in parentheses. Alignments are available upon request.

<table>
<thead>
<tr>
<th>Protein</th>
<th>MUSCLE length</th>
<th>MAFFT length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hippo (Hpo)</td>
<td>676</td>
<td>682</td>
</tr>
<tr>
<td>Salvador (Sav)</td>
<td>763</td>
<td>737</td>
</tr>
<tr>
<td>Warts (Wts)</td>
<td>2637</td>
<td>3366</td>
</tr>
<tr>
<td>Mats (Mats)</td>
<td>227</td>
<td>227</td>
</tr>
<tr>
<td>Yorkie (Yki)</td>
<td>732</td>
<td>787</td>
</tr>
<tr>
<td>Scalloped (Sd)</td>
<td>583</td>
<td>639</td>
</tr>
<tr>
<td>Kibra</td>
<td>1526</td>
<td>1556</td>
</tr>
<tr>
<td>Expanded (Ex)</td>
<td>648 (1495)</td>
<td>648 (1536)</td>
</tr>
<tr>
<td>Merlin (Mer)</td>
<td>737</td>
<td>739</td>
</tr>
</tbody>
</table>