Fungal Smn and Spf30 homologues are mainly present in filamentous fungi and genomes with many introns: Implications for Spinal Muscular Atrophy

Pablo Mier and Antonio J. Pérez-Pulido*

Centro Andaluz de Biología del Desarrollo (CABD), CSIC-UPO. Facultad de Ciencias Experimentales (Área de Genética), Universidad Pablo de Olavide, 41013 Sevilla, Spain

* Corresponding author (ajperez@upo.es, +34 954348652)

Abstract

Spinal muscular atrophy is an important rare genetic disease characterized by the loss of motor neurons, where the main gene responsible is s*mn1*. Orthologous genes have only been characterized in a single fungal genome: *Schizosaccharomyces pombe*. We have searched for putative SMN orthologues in publically available fungal genomes, finding that they are predominately present in filamentous fungi. SMN binding partners and the SPF30 SMN paralogue, which are all involved in mRNA splicing, were found to be present in a similar but non-identical subset of fungal genomes. The *Saccharomycces cerevisiae* yeast genome contains neither smn1 orthologues nor paralogues and it has been suggested that this might be related to the low number of introns in this yeast. Here we have tested this hypothesis by looking at other fungal genomes. Significantly, we find that fungal genomes with high numbers of introns also possess an SMN orthologue or at least its paralogue, SPF30.

Abbreviations: SMA, spinal muscular atrophy; Smn, survival motor neuron; snRNP, small nuclear ribonucleoprotein; hnRNP, heterogeneous nuclear ribonucleoprotein; AC, Accession number; EST, Expressed Sequence Tag; NIG, number of introns per gene; ORF, open reading frame

Keywords: Spinal muscular atrophy; Fungi; Introns; Phylogenetics; Smn; Spf30

1. Introduction

Spinal muscular atrophy (SMA) is a genetic disease characterized by the loss of lower motor neurons. The majority of patients present autosomal recessive inheritance with proximal manifestation of muscle weakness and atrophy [1]. SMA is the most frequent genetic cause of infant deaths and has an incidence of about one in 10,000 live births with a carrier frequency of one in 50 [2].

SMA is caused by the lack or the mutation of the *Smn1* gene (survival motor neuron) [3] whose protein (SMN; UniProt:Q16637) is involved in pre-mRNA splicing and has a specific but poorly understood function in motor neuron axons (see review in [4]). SMN is part of a complex involved in the assembly of small nuclear ribonucleoproteins (snRNP) in the cytoplasm where it exists as an oligomer, but it also operates in nuclear Cajal bodies, where the spliceosome responsible of pre-mRNA splicing is located [5].

SMN possesses several known domains and motifs (sequence patterns), mainly related to binding sites. The N-terminus contains an important region for binding to Gemin2 (also called SIP1), a protein component of the nuclear SMN complex [6]. SMN also possesses a Tudor domain, which has affinity to Sm ribonucleoproteins that form a ring involved in the splicing process [7] and a region with a high frequency of Proline amino acids that is important for binding to proteins such as Profilins [8]. At the C-terminus is the Sm protein binding region and a region responsible for binding Syncrip, a heterogeneous nuclear ribonucleoprotein (hnRNP) implicated in mRNA processing [9].

A partial homologue of SMN, known as SPF30, has also been characterized [10]. This protein has also a Tudor domain and it is necessary for spliceosome assembly, but remains to be shown if it is functionally related to SMN. It has been proposed that SMN and SPF30 could have antagonist functions on cell apoptosis. Talbot and co-workers speculated that SPF30 might have a role in motor neuron development, but this remains to be tested. SPF30 is expressed at high levels in muscle, whose development is closely associated with that of motor neurons.

SMN orthologues have been found in all sequenced vertebrate and invertebrate animals, and several of these have been developed into disease models [11]. Apart from metazoans, a *Smn1* orthologue has also been identified and characterized in fission yeast *Schizosaccharomyces pombe* [12,13,14] which has an identity of approximately 25% at the amino acid level versus SMN. In previous works with this yeast, Smn has been shown to be essential for viability, and it interacts with human SMN and Sm proteins, indicating a remarkable conservation of Smn functional domains in *S. pombe*.

S. pombe Smn is shorter than human SMN but it conserves important sequence patterns, including the N- and C-termini regions but not the Tudor domain. This yeast also possesses the Sip1 and Sm proteins, as well as the Spf30 orthologue, which together have important roles in the splicing process.

Taking in account the mainly constitutive function of SMN and its sequence conservation, the use of fungal genomes to study its function and relationship with human disease has the potential to uncover novel insights into how impaired SMN function can lead to disease.

Interestingly, Smn is not present in *Saccharomyces cerevisiae* and it was proposed that this might be related to the low number of introns present in this yeast [13]. The presence of Smn has not been reported in other fungal genomes but there is likely to be considerable heterogeneity in its conservation given the difference between the two fungal genomes studied so far.

In this work, we set out to test if the presence of Smn in a specific fungal genome is related to a high average number of introns per gene. To accomplish this aim, here we search for Smn orthologues in all the complete fungal genomes available, and find that Smn homologues are only present in a subset of fungal genomes, with most of them belonging to filamentous fungi of the *Pezizomycotina* subphylum. Extending this analysis, we propose several new fungal Smn proteins and binding partners, and finally establish the relationship between the presence Smn homologues and the number of introns in fungal genomes.

2. Materials and methods

2.1. Obtaining fungal genomes

Twenty-six genomes (.fna), proteomes (.faa) and GenBank entries (.gbk) from fungal organisms were downloaded from the National Center for Biotechnological Information (NCBI) FTP server, which is hosted at ftp://ftp.ncbi.nlm.nih.gov/genomes/Fungi/ (14/02/2011). Three genomes were discarded due to the lack of significant number of genes and/or proteins. The selected fungal species were: *Aspergillus fumigatus*, *Aspergillus nidulans*, *Aspergillus niger*, *Candida dubliniensis*, *Candida glabrata*, *Cryptococcus gattii*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Encephalitozoon intestinalis*, *Eremothecium gossypii*, *Gibberella zeae*, *Kluyveromyces lactis*, *Lachancea thermotolerans*, *Magnaporthe grisea*, *Neurospora crassa*, *Pichia pastoris*, *Pichia stipitis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Ustilago maydis*, *Yarrowia lipolytica*, and *Zygosaccharomyces rouxii*.

2.2. Orthologue search

To look for sequence orthologues (homologues and functionally equivalent proteins in different species), a local two-way protein Blast was performed, where a query sequence is used to search within a proteome, and the best hit is then used as the query sequence for a second Blast against the proteome corresponding to the first sequence. To consider a candidate protein as a true orthologue , the two obtained Blast alignments needed to meet the following criteria: (a) e-value lower than or equal to 5e-05, (b) identity value higher than or equal to 25%, (c) sequence query length at least 80% covered by the alignment, and (d) a similar length for both the protein and the putative orthologue. Blast analyses were performed using the following parameters: Blosum62 matrix, 11 as cost to open a gap, 1 as cost to extend a gap, and no low-complexity filter.

When no candidate orthologue was found, a genome search, the more sensitivity PSI-Blast [15], and Hmmer [16] tools were used, but no additional results were found in this way.

2.3. Databases

Initial protein and gene sequence data, as well as data for building the phylogeny of the fungi under study was obtained from NCBI RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/). Resulting candidate protein sequences were then used to search the UniProt Knowledgebase (http://www.uniprot.org) by blast in order to take advantage of the wealth of protein annotation available in this database. Accession numbers (AC) were collected and additional information about each protein was extracted from this database. Finally, Expressed Sequence Tags (EST) for studied organisms were downloaded from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) and a simple Tblastn was carried out to validate the initial sequence selection.

2.4. Calculation of the number of introns per gene (NIG)

The number of introns per gene (NIG) parameter was calculated using our own program written in the Perl programming language. The program searches the CDS field for all the genes in a fungal genome and computes the total number of commas within it. This represents the number of introns in a given gene.

e.g.: CDS *join(30280..30343,30401..30489,30540..30636,30664..30782)* => this gene have 3 introns.

Finally, the total number of introns in a genome is divided by the total number of genes to obtain the NIG value for each genome. The calculated NIG have been compared with that published for several fungi and the results are similar [17,18].

3. Results

3.1. Search for Smn gene and protein in fungal genomes

With the aim of identifying new Smn sequences in fungal species, the *S. pombe* Smn protein sequence was used as the query sequence in a two-way Blast strategy against complete fungal proteomes, complemented by other bioinformatics approaches. In this way, five putative Smn orthologues were found, with an average similarity of 28% relative to the query sequence, in the following species: *A. fumigatus*, *A. nidulans*, *A. niger*, *N. crassa* and *Y. lipolytica* (see Table 1 for detailed information of the genes identified in this work).

All of the protein orthologues identified, including the *S. pombe* sequence, have short conserved patterns at their N- and C-terminal ends, with the remaining sequence weakly conserved (Fig. 1a). Part of the Sip1 protein binding region (with the WDDxxL motif) is conserved within the N-terminus (Fig. 1b). The C-terminus Sm ribonucleoprotein binding and oligomerization domain contains another conserved sequence motif in fungal Smn orthologues corresponding to a WY box. The central region in vertebrate orthologues mainly corresponds to a Tudor domain involved in binding to methylarginine-containing proteins [7]. The Tudor domain is not conserved between vertebrate and fungi Smn orthologues.

Other fungal Smn proteins were found using alternative *in silico* methods as explained in the materials and methods, mainly supported by searches in the EST database.

3.1.1. The putative N. crassa Smn sequence may be incorrectly annotated

When we searched for Smn in the *N. crassa* proteome we found a protein of 229 amino acids (GI:164428129). In addition to the normally conserved N-terminus region, this sequence has an extended N-terminal region, which does not globally align with the other Smn sequences. Two different, unreviewed sequences for this protein exist in the UniProt database: one of them is the same as the proteome sequence (Q7RY22), and the other is a truncated version of it (Q96UC1). Both UniProt sequences align closely in the C-terminal region but not in the N-terminal. However, when we ran a Tblastn against the *N. crassa* EST database using *S. pombe* Smn as query sequence, we identified an alternative sequence which appropriately aligns at both terminal regions and with barely any gaps. Thus, we conclude that the EST translation sequence represents the most likely Smn homologue candidate sequence in *N. crassa* (Fig. 1a) which is 73 amino acids shorter than the sequence present in the current protein databases.

3.1.2. The putative M. grisea Smn candidate was identified from an EST sequence

As our initial search for Smn homologues in the *M. grisea* proteome was unsuccessful, we next searched the EST database using Tblastn with *S. pombe* Smn as query sequence. An EST sequence coding for a putative protein with the conserved N-terminal region was initially found (GenBank:DC972816.1). The corresponding amino acid sequence was used as the query in a UniProt Blast identifying another unreviewed protein, possessing both N-terminal and C-terminal conserved regions, and thus represents the putative *M. grisea* Smn homologue.

3.1.3. A point deletion is present in the sequenced G. zeae genome

The search for Smn in the *G. zeae* proteome was also unsuccessful. However, there is a computationally predicted protein in this proteome which contains the N-terminal motif corresponding to Smn (GI:46117098). Since this pattern is uncommon in proteins, the sequence might correspond to a Smn orthologue, albeit lacking the conserved C-terminus. We speculated that the absence of this region might result from a frame shift in the theoretical protein sequence caused by a small error in the *G. zeae* genome sequence. To address this possibility, we calculated the protein sequence for all 6 possible open reading frames (ORF) for the corresponding DNA sequence (coding sequence; GI:46117097) and then searched for sequences corresponding to the conserved C-terminal in all the ORFs other than original coding one (+1). Significantly, we found a region of ORF +3 which showed high conservation with the C-terminal Smn region. This strongly suggests that the genome sequence in this region (between N-terminal and C-terminal patterns) contains a deletion, and a sequence with all the characteristics of other fungal Smn can be found if an undefined nucleotide (N) is inserted within it (see sequence in Fig. 1a, where "X" marks an undefined amino acid introduced by this nucleotide).

As can be seen in the Fig. 1a, all proposed Smn proteins conserve both N-terminus and C-terminal regions, and the molecular phylogeny using these sequences is consistent with the evolutionary relationship between the corresponding organisms (Fig. 2).

3.2. Sip1 and Sm proteins are associated with Smn in metazoans but only Sm proteins are present in fungi

Human SIP1 (also called Gemin2, and Yip11 in *S. pombe*) is part of the SMN complex. This complex is distributed between the nucleus and cytoplasm of human cells, while it predominantly accumulates in the nucleus of *S. pombe* [12]. The human SMN complex plays an essential role in spliceosomal snRNP assembly in the cytoplasm and is required for pre-mRNA splicing in the nucleus. SIP1 has been shown to bind to the N-terminus of SMN [6].

To explore the relationship between Sip1 and Smn in fungi, were searched for Sip1 orthologues in the same way as we did for Smn. Surprisingly, Sip1 was generally found in genomes which lack clear Smn homologues except for *S. pombe* and *Y. lipolytica* which, remarkably, are the only yeasts that possess Smn proteins (Fig. 3).

On the other hand, SMN helps in the assembly of spliceosomal snRNPs that are composed of seven Sm proteins (B/B’, D1, D2, D3, E, F and G) forming a seven-member ring core structure that encircles RNA in the nuclear splicing process.

We also searched fungal genomes for orthologues of the Smn-complex Sm proteins. The C-terminus of Smn contains the motif primarily responsible for binding Sm proteins although a role for the non-conserved central region has also been proposed [19]. Almost all the genomes analyzed contain orthologues corresponding to all the Sm proteins (data not shown), showing that the spliceosome complex is conserved in all fungal genomes.

3.3. Spf30 is a Smn paralogue that usually appears in the fungi which have Smn

Vertebrate Smn has a partial paralogue of similar length named SPF30, which shares the Tudor domain and Sm binding site [10,20]. SPF30 was identified as a constituent of the spliceosome complex and it localizes in the nucleus. SPF30 localizes with SMN protein at Cajal bodies, a site of snRNP accumulation and Sm subcomplex assembly [21]. However, the specific functional relationship between SMN and SPF30 remains unclear.

To test the evolutionary relationship between Smn and Spf30 in fungi, orthologues of *S. pombe* Spf30 were searched for in fungal proteomes. As expected we found that this protein mainly appears in species where Smn is also present (Fig. 3). Interestingly, we found some exceptions, such as in fungi from the *Basidiomycota* phylum *Cryptococcus* and *U. maydis* where Spf30 was present but Smn was not.

3.4. Fungi with a high rate of splicing have Smn and/or Spf30

When *S. pombe* Smn was discovered it was proposed that *S. cerevisiae* lacked the Smn protein because of the low number of introns of this budding yeast [13]. Our results confirm the absence of Smn from the *S. cerevisiae* genome. Moreover, our identification of putative Smn orthologues in several other fungal species allow us to test the hypothesis linking Smn and intron number. To this end, the number of introns per gene (NIG) in all sequenced fungal genome was calculated. Each genome was assigned a number indicating the ratio between the total number of introns from all the genes and the total number of genes. Our results range from fungal genomes with low NIG values, such as *S. cerevisiae*, to genomes with moderate to high NIG values, such as *S. pombe* (Fig. 3). We observed a clear correlation between high NIG values and the presence of Smn in the genomes tested. However, this did not hold true for *Basidiomycota* genomes since they have higher NIG values but lack Smn protein homologues. However, *Cryptococcus* and *U. maydis* have the Smn paralogue, Spf30, suggesting that it could play an important role in the splicing process, perhaps complementing or replacing Smn function in this large phylum.

4. Discussion

This work presents a detailed bioinformatics search within fungal genomes for a protein related to an important human disease. The results have revealed the evolutionary conservation of the *Smn* gene and its coding protein in fungi. The Smn protein is conserved in a series of fungi which are mainly grouped in *Pezizomycotina* including filamentous ascomycetes (Fig. 4). The only yeast fungi with Smn are the very well studied *S. pombe*, and *Y. lipolytica*. The *Pezizomycotina* subphylum where Smn is predominately found is formed by fungi presenting mycelia with thread-like hyphae. These hyphae are analogous to neural axons, and even share some common elements with them, such as the microtubules involved in growth and vesicle transport [22]. The link between a lack of Smn and SMA disease is probably related to the recently demonstrated role for Smn in mRNA transport along these axons [23]. Thus, these fungi have the potential to be key model organisms for understanding the molecular mechanisms that link impaired Smn function to SMA. Laboratory experiments showing the location and relation of both Smn and Spf30 in fungal hyphae may help to understand how these proteins interact with the cell prolongations and what their function there. These findings could then be extrapolated to neural axons. Thus, our progress towards understanding of SMA disease could be accelerated due to the advantages of working with fungi.

Interestingly, the candidate Smn homologues identified in this work have two highly conserved motifs, one at the N-terminus with the consensus WDDxxL and the other at the C-terminus with a WY box. In addition, all of the putative fungal Smn homologues, but not the *S. pombe* one, possess a similar additional N-terminus pattern with the consensus W[DN][Ek] (Fig. 1a), which might have arisen from a duplication event.

Here we have shown that species that have a high NIG value possess Smn or at least its paralogue Spf30. In most species this protein is normally present together with Smn, with the exception of the *Basidiomycota* phylum where Spf30 is present but Smn is not. All of the putative Spf30 sequences identified in this work conserve the C-terminal motif and the Tudor domain, consistent with previous findings [24]. In human, both SMN and SPF30 proteins are localized within the nucleus to Cajal bodies, but only SPF30 is also found in nuclear speckles together with other splicing factors, whereas SMN is additionally found in the cytoplasm [20]. Furthermore, SMN does not complement the lack of SPF30 [20]. Thus, the SMN paralogue, SPF30, could complement but not replace the function of SMN. But in fungi, we find that Spf30 sequences conserve the Sm protein binding site present in Smn, although binding between Spf30 and Sm has not been shown. Strikingly, even though metazoan Smn has a Tudor domain, fungal Smn does not, suggesting that Smn and Spf30 might have complementary functions in fungi. If this were true, fungal Spf30 would be responsible for Sm protein binding thanks to its Tudor domain, and fungal Smn would have the remaining functions found for this protein in metazoans.

The presence or absence of Smn in a organism has previously been proposed to be related to active intron processing. In this work, we have shown that the presence of Smn is directly associated to the number of introns per gene in a given fungal genome. Despite the presence of Smn in different and heterogeneous groups of fungi, all of them have a high NIG value in common. Only the *Basidiomycota* phylum which have a high NIG value but lacks Smn. Interestingly these genomes contain the Smn paralogue Spf30, again suggesting that Spf30 could be providing Smn function in this group of fungi. Such a situation could be a common one, given that the constitutive functions of such proteins in fungal organisms are likely to have subtle differences relative to metazoan organisms. In fact, despite the expected importance of Smn, the majority of yeast lack it. However, many more complete genomes will need to be looked at verify these hypotheses.

Some Smn orthologues were more difficult to identify, highlighting the complexity and low global conservation of this protein, possessing only two conserved short motifs. These problems made it necessary to develop new integrative bioinformatics protocols to search for orthologous sequences with these characteristics. The *in silico* procedure used in this work has allowed us to identify Smn orthologues in fungal organisms where the protein was initially thought to be absent, or its gene structure erroneously annotated. This fact is mainly due to the low conservation of the central region of Smn. In metazoans, this region contains the Tudor domain involved in Sm protein binding, which has not been found in fungi. Moreover, in metazoans this region contains a long stretch of sequence with a compositional bias towards poly-proline that also does not appear in fungi (Fig. 1b). Recently, in an elegant *in silico* analysis, the compositional bias of regions located towards the extremities of proteins has been linked to new evolutionary elements associated with proteins with high numbers of binding partners [25]. In fact, this region has been shown to be important for binding to the Profilin [8], an actin-binding protein that affects the structure of the cytoskeleton, and is highly expressed in the spinal cord. This low complexity region is much wider in vertebrates than invertebrates and it is completely absent from fungal Smn orthologues. Remarkably, the putative *U. maydis* Spf30 orthologues have a clear poly-proline region within its C-terminus (data not shown), which is similar to vertebrate Smn regions. It is also interesting to note that fungal genomes also contain the profilin protein, which all together suggests another complementary function for Smn and Spf30 in fungal genomes. This putative complementarity is also suggested because when we compare fungal Smn and Spf30 the similarity is very low (data not show). But when human SMN and SPF30 are aligned, the similarity is high, especially within the Tudor domain. Since fungal Spf30 has a recognizable Tudor domain, and this region is not similar to the central region of Smn, we can confirm that Smn lacks a Tudor domain and it could be complemented by Spf30.

We also searched for orthologues of the Smn binding protein Sip1, which is important for the splicing related function of Smn. Surprisingly, it is only present in some yeasts but not in fungi that possess Smn, apart from *S. pombe* and *Y. lipolytica*. This could be due to hypothetical Sip1 homologues, which conserve subtle sequence properties that remain undetected by sequence analysis. For example, it has been proposed that *S. cerevisiae* has a Sip1 orthologue that is very different from *S. pombe* Sip1 [6] called Brr1 (UniProt: Q99177). Even the other Sip1 orthologues found in this work exhibit very low sequence conservation (see Table 1). Furthermore, when *S. cerevisiae* Brr1 is compared to the Sip1 sequences proposed here, all of them conserve important amino acids throughout the alignment (data not shown). These observations suggest that Sip1 might have Smn-independent functions in fungi, especially in yeasts.

In summary, we have confirmed a relationship between the number of introns per gene and the presence of Smn, except for *Basidiomycota* genomes, where Spf30 is present but Smn is not, which could mark the point where Smn arose in evolution. It could initially have arisen as a splicing factor, that later it would acquired new functions related to mRNA transport and others activities related to the SMA phenotype in the animal motor nervous system.

S. pombe is one of the most widely studied fungal model organisms, but it is not necessarily the most appropriate for studying a particular process. The data we have presented here suggest that other fungal species may also represent valuable model organisms for the study of important human diseases. Understanding the conservation and evolution of different genes related to diseases such as SMA should make it easier to select model organisms where the underlying processes are most likely to be similar to those occurring in human cells. This should allow researchers to obtain better and more relevant data that can then be applied to our understanding of human diseases.

Acknowledgments

We thank Manuel Corredera and Carlos M. Guevara for helping us get started on this issue, and John Pearson for his critical reading of the manuscript. We also acknowledge institutional support from the Junta de Andalucía to the CABD.

References

[1] B. Wirth, L. Brichta, E. Hahnen, Spinal muscular atrophy: from gene to therapy, Semin. Pediatr. Neurol. 12 (2006) 121-131.
[2] M.R. Lunn, C.H. Wang, Spinal muscular atrophy, Lancet 371 (2008) 2120-2133.
[3] S. Lefebvre, L. Bürglen, S. Reboullet, O. Clermont, P. Burlet, L. Viollet, B. Benichou, C. Cruaud, P. Millasseau, M. Zeviani, Identification and characterization of a spinal muscular atrophy-determining gene, Cell 80 (1995) 155-165.
[4] A.H.M. Burghes, C.E. Beattie, Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick?, Nat. Rev. Neurosci. 10 (2009) 597-609.
[5] S.J. Kolb, D.J. Battle, G. Dreyfuss, Molecular Functions of the SMN Complex, J. Child Neurol. 22 (2007) 990-994.
[6] Q. Liu, U. Fischer, F. Wang, G. Dreyfuss, The Spinal Muscular Atrophy Disease Gene Product, SMN, and Its Associated Protein SIP1 Are in a Complex with Spliceosomal snRNP Proteins, Cell 90 (1997) 1013-1021.
[7] J. Côté, S. Richard, Tudor Domains Bind Symmetrical Dimethylated Arginines, J. Biol. Chem. 280 (2005) 28476 -28483.
[8] T. Giesemann, S. Rathke-Hartlieb, M. Rothkegel, M, J.W. Bartsch, S. Buchmeier, B.M. Jockusch, H.A. Jockusch, Role for Polyproline Motifs in the Spinal Muscular Atrophy Protein SMN, J. Biol. Chem. 274 (1999) 37908 -37914.
[9] Z. Mourelatos, L. Abel, J. Yong, N. Kataoka, G. Dreyfuss, SMN interacts with a novel family of hnRNP and spliceosomal proteins. EMBO J. 20 (2001) 5443-5452.
[10] K. Talbot, I. Miguel-Aliaga, P. Mohaghegh, C.P. Ponting, K.E. Davies, Characterization of a Gene Encoding Survival Motor Neuron (Smn)-Related Protein, a Constituent of the Spliceosome Complex, Hum. Mol. Genet. 7 (1998) 2149-2156.
[11] A. Schmid, C.J. DiDonato, Animal models of spinal muscular atrophy, J. Child Neurol. 22 (2007) 1004-1012.
[12] S. Hannus, D. Buhler, M. Romano, B. Seraphin, U. Fischer, The Schizosaccharomyces pombe protein Yab8p and a novel factor, Yip1p, share structural and functional similarity with the spinal muscular atrophy-associated proteins SMN and SIP1, Hum. Mol. Genet. 9 (2000) 663-674.
[13] N. Owen, C.L. Doe, J. Mellor, K.E., Characterization of the Schizosaccharomyces pombe orthologue of the human survival motor neuron (SMN) protein, Hum. Mol. Genet. 9 (2000) 675-684.
[14] S. Paushkin, B. Charroux, L. Abel, R.A. Perkinson, L. Pellizzoni, G. Dreyfuss, The Survival Motor Neuron Protein of Schizosacharomyces pombe, J. Biol. Chem. 275 (2000) 23841-23846.
[15] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389-402.
[16] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res 39 (2011) W29-W37.
[17] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann et al., Life with 6000 genes. Science 274:546 (1996) 563-567.
[18] V. Wood, R. Gwilliam, M.A. Rajandream, M. Lyne, R. Lyne, A. Stewart et al., The genome sequence of Schizosaccharomyces pombe. Nature 415 (2002) 871-880.
[19] D. Bühler, V. Raker, R. Lührmann, U. Fischer, Essential Role for the Tudor Domain of SMN

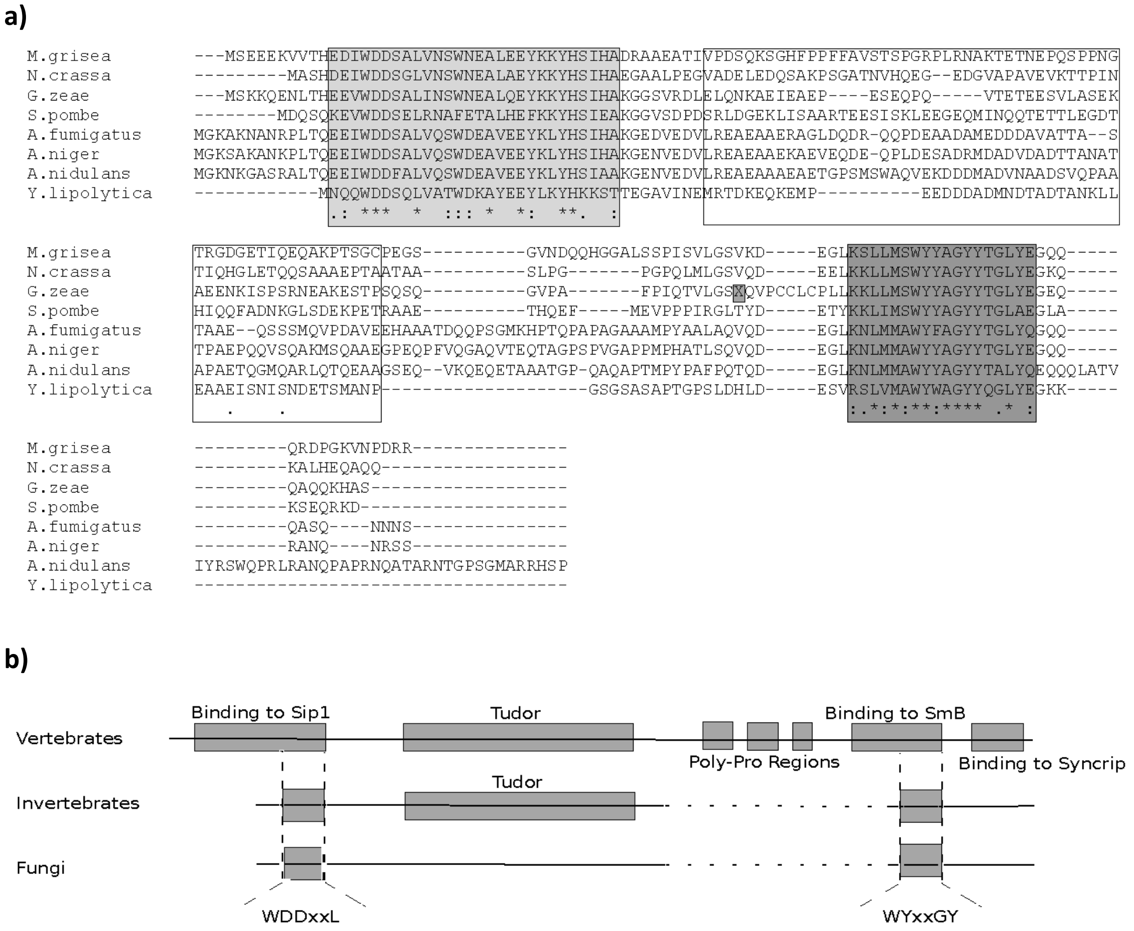
in Spliceosomal U snRNP Assembly: Implications for Spinal Muscular Atrophy, Hum. Mol. Genet. 8 (1999) 2351-2357.
[20] J. Rappsilber, P. Ajuh, A.I. Lamond, M. Mann, SPF30 Is an Essential Human Splicing Factor Required for Assembly of the U4/U5/U6 Tri-small Nuclear Ribonucleoprotein into the Spliceosome, J. Biol. Chem. 276 (2001).
[21] J.T. Little, M.S. Jurica, Splicing Factor SPF30 Bridges an Interaction between the Prespliceosome Protein U2AF35 and Tri-small Nuclear Ribonucleoprotein Protein hPrp3, J. Biol. Chem. 283 (2008) 8145 -8152.
[22] I. Schuchardt, D. Aßmann, E. Thines, C. Schuberth, G. Steinberg, Myosin-V, Kinesin-1, and Kinesin-3 Cooperate in Hyphal Growth of the Fungus Ustilago maydis, Mol. Biol. Cell 16 (2005) 5191-5201.
[23] W. Rossoll, S. Jablonka, C. Andreassi, A.K. Kröning, K. Karle, U.R. Monani, M. Sendtner, Smn, the spinal muscular atrophy–determining gene product, modulates axon growth and localization of β-actin mRNA in growth cones of motoneurons, J. Cell Biol. 163 (2003) 801-812.
[24] G. Meister, S. Hannus, O. Plottner, T. Baars, E. Hartmann, S. Fakan, B. Laggerbauer, U. Fischer, SMNrp is an essential pre-mRNA splicing factor required for the formation of the mature spliceosome, EMBO J. 20 (2001) 2304-2314.
[25] A. Coletta, J. Pinney, D. Solis, J. Marsh, S. Pettifer, T. Attwood, Low-complexity regions within protein sequences have position-dependent roles, BMC Syst. Biol. 4 (2010) 43.

Table 1. Identifiers and Blast output parameters for putative Smn, Spf30, and Sip1 fungal protein sequences. The identifier is the Accession Number from UniProt or RefSeq, the identity percentage refers to a global alignment versus *S. pombe* sequence (local alignments provide somewhat higher values), and the e-value refers to local alignments found by Blast.

Gene	Organism	Identifier	%Identity	E-value	Length (aas.)
Smn	*S.pombe*	Q09808	100%	0	152
	A.nidulans	C8V942	25%	1e-10	206
	*G.zeae**	A2QHQ5	30%	5e-12	147
	A.niger	XP_384567	28%	6e-07	152
	*M.grisea**	A4QYJ6	26%	4e-05	167
	A.fumigatus	Q4WNN4	27%	5e-11	171
	*N.crassa**	Q7RY22	33%	5e-11	149
	Y.lipolytica	Q6C798	27%	7e-07	126
Spf30	*S.pombe*	O94519	100%	0	311
	C.neoformans	Q5KKF2	22%	5e-10	229
	C.gatti	E6R3I4	22%	8e-08	250
	*A.nidulans**	C8VBJ7	31%	1e-13	289
	A.niger	A2QX86	30%	1e-19	291
	G.zeae	XP_388559	27%	2e-09	279
	A.fumigatus	Q4WMG8	29%	2e-17	297
	N.crassa	Q7S6P1	26%	3e-15	370
	*U. maydis***	Q4P543	20%	2e-05	247
Sip1	*S.pombe*	Q9P347	100%	0	235
	P.stipitis	A3LUZ7	20%	3e-06	228
	Y.lipolytica	Q6CDV1	20%	3e-08	262
	P.pastoris	C4QHZ1	22%	4e-08	261
	D.hansenii	Q6BL84	20%	2e-05	289
	K.lactis	Q6CVY1	21%	2e-05	326

[*] Candidate sequences do not exactly correspond with the identifier shown, due to manual sequence revision (see results).

[**] It was found by Blast using *C. neoformans* Spf30 as the query sequence.



```
M.grisea        ---MSEEEKVVTHEDIWDDSALVNSWNEALEEYKKYHSIHADRAAEATIVPDSQKSGHFPPFFAVSTSPGRPLRNAKTETNEPQSPPNG
N.crassa        ---------MASHDEIWDDSGLVNSWNEALAEYKKYHSIHAEGAALPEGVADELEDQSAKPSGATNVHQEG--EDGVAPAVEVKTTPIN
G.zeae          ---MSKKQENLTHEEVWDDSALINSWNEALQEYKKYHSIHAKGGSVRDLELQNKAEIEAEP----ESEQPQ-----VTETEESVLASEK
S.pombe         --------MDQSQKEVWDDSELRNAFETALHEFKKYHSIEAKGGVSDPDSRLDGEKLISAARTEESISKLEEGEQMINQQTETTLEGDT
A.fumigatus     MGKAKNANRPLTQEEIWDDSALVQSWDEAVEEYKLYHSIHAKGEDVEDVLREAEAAERAGLDQDR-QQPDEAADAMEDDDAVATTA--S
A.niger         MGKSAKANKPLTQEEIWDDSALVQSWDEAVEEYKLYHSIHAKGENVEDVLREAEAAEKAEVEQDE-QPLDESADRMDADVDADTTANAT
A.nidulans      MGKNKGASRALTQEEIWDDFALVQSWDEAVEEYKLYHSIAAKGENVEDVLREAEAAAEAETGPSMSWAQVEKDDDMADVNAADSVQPAA
Y.lipolytica    ------------MNQQWDDSQLVATWDKAYEEYLKYHKKSTTEGAVINEMRTDKEQKEMP----------EEDDDADMNDTADTANKLL
                             .: ***  *  ::: *  *:  **.  :

M.grisea        TRGDGETIQEQAKPTSGCPEGS----------GVNDQQHGGALSSPISVLGSVKD-----EGLKSLLMSWYYAGYYTGLYEGQQ-----
N.crassa        TIQHGLETQQSAAAEPTAATAA----------SLPG-------PGPQLMLGSVQD-----EELKKLLMSWYYAGYYTGLYEGKQ-----
G.zeae          AEENKISPSRNEAKESTPSQSQ----------GVPA-------FPIQTVLGSXQVPCCLCPLLKKLLMSWYYAGYYTGLYEGEQ-----
S.pombe         HIQQFADNKGLSDEKPETRAAE----------THQEF-----MEVPPPIRGLTYD-----ETYKKLIMSWYYAGYYTGLAEGLA-----
A.fumigatus     TAAE--QSSSMQVPDAVEEHAAATDQQPSGMKHPTQPAPAGAAAMPYAALAQVQD-----EGLKNLMMAWYFAGYYTGLYQGQQ-----
A.niger         TPAEPQQVSQAKMSQAAEGPEQPFVQGAQVTEQTAGPSPVGAPPMPHATLSQVQD-----EGLKNLMMAWYYAGYYTGLYEGQQ-----
A.nidulans      APAETQGMQARLQTQEAAGSEQ--VKQEQETAAATGP-QAQAPTMPYPAFPQTQD-----EGLKNLMMAWYYAGYYTALYQEQQQLATV
Y.lipolytica    EAAEISNISNDETSMANP--------------------GSGSASAPTGPSLDHLD-----ESVRSLVMAWYWAGYYQGLYEGKK-----
                   .    .                                                      :.*:*:**:**** .* :

M.grisea        ---------QRDPGKVNPDRR---------------
N.crassa        ---------KALHEQAQQ------------------
G.zeae          ---------QAQQKHAS-------------------
S.pombe         ---------KSEQRKD--------------------
A.fumigatus     ---------QASQ----NNNS---------------
A.niger         ---------RANQ----NRSS---------------
A.nidulans      IYRSWQPRLRANQPAPRNQATARNTGPSGMARRHSP
Y.lipolytica    ------------------------------------
```

Fig. 1. Smn conservation. (a) Multiple alignment of known and proposed Smn fungal proteins built using Clustalw. The N-terminal and C-terminal conserved region are highlighted in a grey box, and the region where the Tudor domain should be is highlighted with a transparent box. An 'X' with a gray background appears in the position where a 'N' was inserted in the DNA sequence of *G. zeae*. The asterisks mark identical amino acid positions, and points and colons mark similar positions. (b) Smn domains and motifs extracted from vertebrate protein sequences in UniProt, and the corresponding sequences in invertebrates and fungi. Amino acids in N- and C- terminal regions conserved between all the organisms are highlighted, and the absence of a region in a organism group is marked with a dotted line.

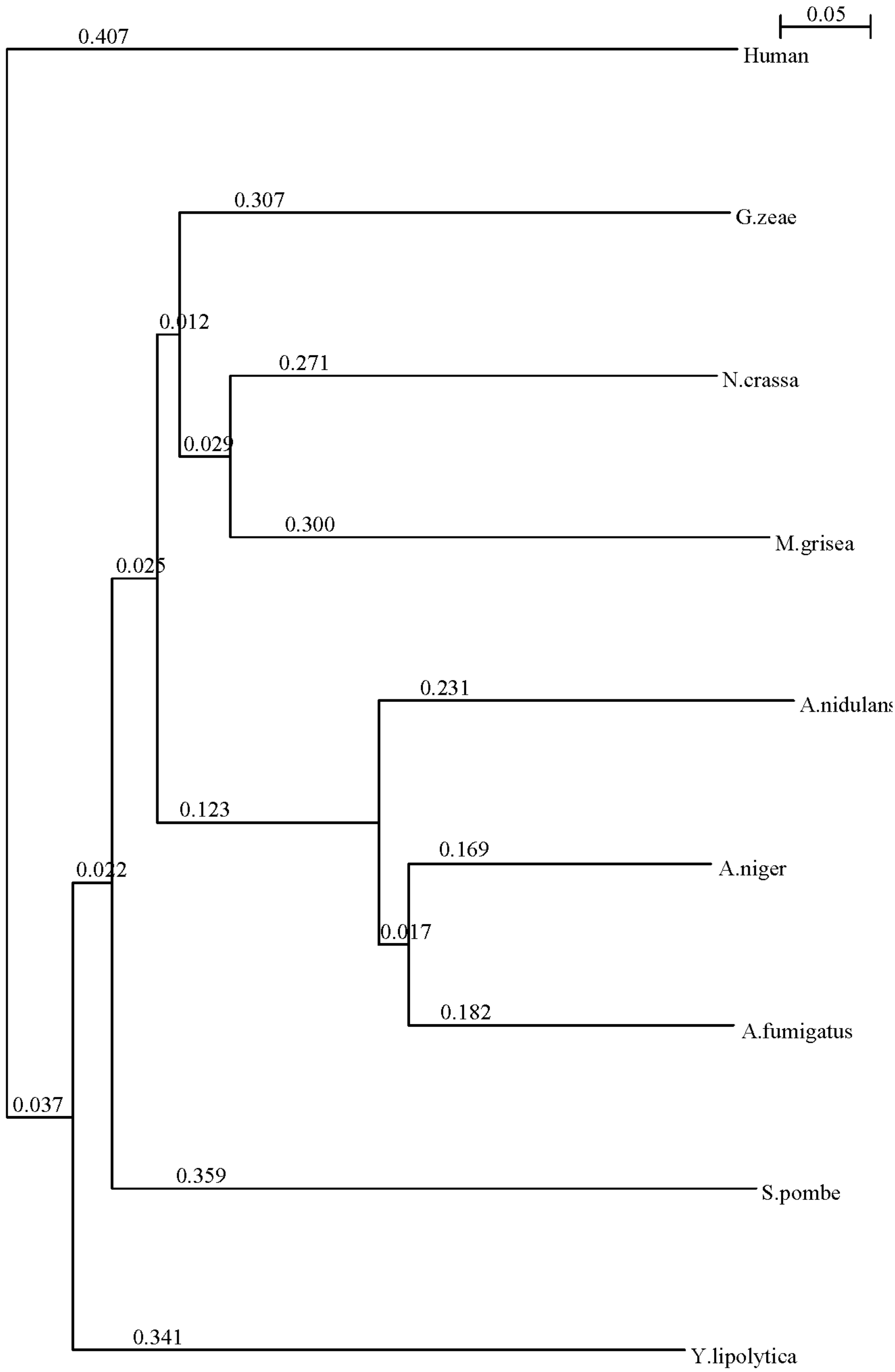


Fig. 2. Molecular phylogeny from the fungal Smn sequences proposed in this work. The sequences of Fig. 1a and the neighbor-joining algorithm were used. Branch lengths are shown. The human SMN was used as the outgroup to root the tree.

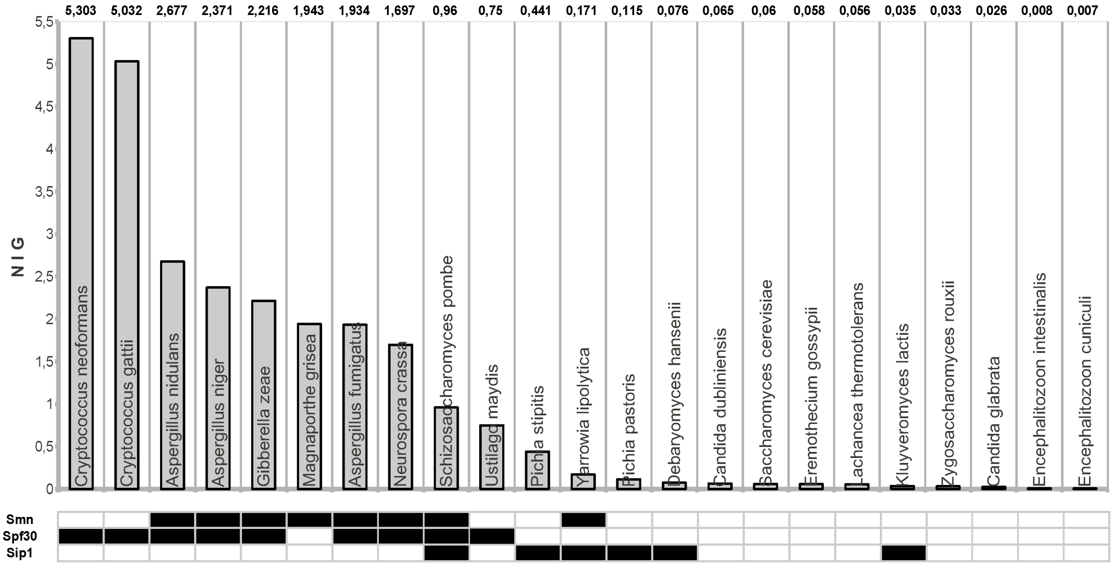


Fig. 3. Protein orthologues identified and number of introns per genes (NIG) for fungal genomes. Filled cells in the lower table indicate that the protein is found in a specific proteome (the name of each organism appears in the graph). The upper histogram shows the NIG value calculated for each fungal genome. Genomes have been ordered from higher to lower values, and the exact NIG is given above. The table with the presence/absence can be compared with the upper NIG values.

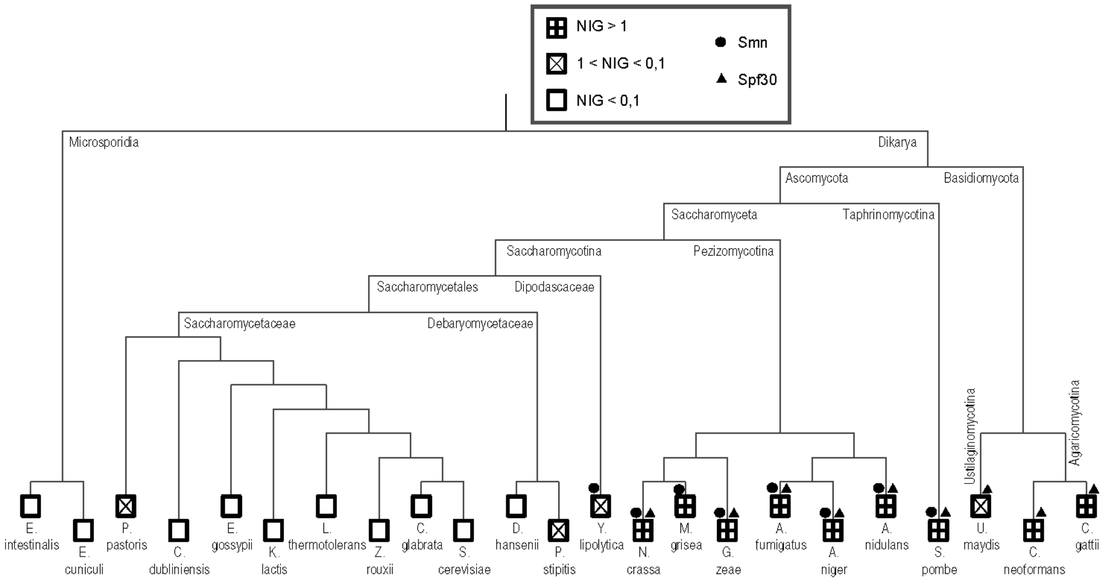


Fig. 4. Phylogenetic relationships between fungal organisms together with NIG values and the presence of Smn and/or Spf30. The NIG is grouped into three classes, and the main taxon for each branch is represented.