Functional insights into the core-TFIIH from a comparative survey

Florence Bedez, Benjamin Linard, Xavier Brochet, Raymond Ripp, Julie D. Thompson, Dino Moras, Odile Lecompte, Olivier Poch

Functional insights into the core-TFIIH from a comparative survey

Florence Bedez, Benjamin Linard, Xavier Brochet, Raymond Ripp, Julie D. Thompson, Dino Moras, Odile Lecompte, Olivier Poch.

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS, INSERM, UDS), BP163, 67404 Illkirch Cedex, France, phone +33 (03) 88 65 32 94, Fax 33 (03) 88 65 32 76, E-mail : florence.bedez@ac-strasbourg.fr.

TFIIH is a eukaryotic complex composed of two subcomplexes, the CAK (Cdk Activating Kinase) and the core-TFIIH. The core-TFIIH, composed of seven subunits (XPB, XPD, P62, P52, P44, P34, P8), plays a crucial role in transcription and repair. Here, we performed an extended sequence analysis to establish the accurate phylogenetic distribution of the core-TFIIH in 63 eukaryotic organisms. In spite of the high conservation of the seven subunits at the sequence and genomic levels, the non enzymatic P8, P34, P52 and P62 are absent from one or a few unicellular species. To gain insight into their respective roles, we undertook a comparative genomic analysis of the whole proteome to identify the gene sets sharing similar presence/absence patterns. While little information was inferred for P8 and P62, our studies confirm the known role of P52 in repair and suggest for the first time the implication of the core TFIIH in mRNA splicing via P34.

Keywords : TFIIH; transcription; repair; splicing, comparative genomic analysis; P34.

1. Introduction

TFIIH is a eukaryotic multiprotein complex initially identified as a General Transcription Factor (GTF) of class II genes. During transcription initiation, TFIIH unwinds DNA through ATPase/helicase activity and promotes the formation of a transcriptionally open complex (Zurita & Merino, 2003). In addition, it specifically phosphorylates the fifth serine (Ser5) of the heptapeptide repeat present in the C-terminal domain (CTD) of RBP1, the largest subunit of the RNA Polymerase II (RNA PolII). Ser5 phosphorylation is thought to facilitate RNA pol II escape from the promoter and the transition from transcription initiation to elongation. It may also serve as a signal for binding of the capping and splicing factors, as well as the histone methyltransferase Set1, to the early elongating RNA PolII (Schroeder *et al.*, 2000; Zurita & Merino, 2003). In contrast to other GTFs, TFIIH is also involved in other vital cellular processes, such as nucleotide excision repair (NER), cell cycle regulation and transcription of ribosomal RNA genes (Zurita & Merino, 2003). Several lines of evidence also suggest that the TFIIH complex may participate in mRNA processing (Damgaard *et al.*, 2008; Hong *et al.*, 2009; Kanin *et al.*, 2007; Viladevall *et al.*, 2009). This functional modularity seems to be related to the highly dynamic composition of TFIIH that has been elegantly observed during early embryo development in *Drosophila* (Aguilar-Fuentes *et al.*, 2006) and more recently during the incision/excision steps of the NER in human (Coin *et al.*, 2008). TFIIH is organized into two major sub-complexes, the core-TFIIH and the CAK (Cdk Activating Kinase) (Table 1).

The functionally diverse CAK subcomplex is composed of the CDK7, CYCLINH and MAT1 proteins and is exclusively found in Eucarya. When associated with the core-TFIIH, it phosphorylates the CTD of RNA polII in all Eucarya. Prokaryotes lack both the CTD and the

CAK. As a free trimeric complex, the CAK regulates the cell-division cycle by phosphorylating various cell cycle cyclin dependant kinases (cdks) except in *Saccharomyces cerevisiae* where these phosphorylations are performed by a monomeric kinase CAK1, very distantly related to CDK7.

The core-TFIIH **c**ontains 7 subunits (Table 1), which are highly conserved between animals, plants and fungi. For the sake of simplicity, the subunits will be named according to the nomenclature of the human core-TFIIH. XPD and XPB, two ATP-dependent helicases, catalyse the unwinding of the DNA duplex at promoters during transcription as well as at DNA lesions during NER (Zurita & Merino, 2003). XPD and XPB homologs have been detected in Prokaryotes, but their function is still poorly understood and seems to be related to NER rather than to transcription (Rouillon & White, 2011). P44 exhibits an ubiquitin ligase activity *in vitro* in *S. cerevisiae* and participates, together with P62, P52 and P34, in protein-protein interactions to maintain the core-TFIIH architecture. In sharp contrast to the other six subunits, P8 is not essential for cell viability and seems to act as an accessory protein in the NER (Ranish *et al.*, 2004). Besides their structural role, P52 and P44 also act as regulatory proteins for the activities of XPD and XPB, respectively (Coin *et al.*, 2007). Currently, little is known about the functional role(s) of P34 and P62. P34 contains a single C-terminus Zinc motif (C4) and has been shown to interact with the Zinc finger domain of P44 through its N-terminal region (Fribourg *et al.*, 2001), whereas P62 is characterized by a N-terminal PH/PTB domain (Gervais *et al.*, 2004) and two folding units, so called BSD domains (Doerks *et al.*, 2002; Jawhari *et al.*, 2004). The PH/PTB domain is known to contact the XPG endonuclease (Gervais *et al.*, 2004) or transcriptional activators (Kwek *et al.*, 2004), whereas the BSD domains are required for core-TFIIH assembly by binding with the P44 subunit (Matsui *et al.*, 1995; Tremeau-Bravard *et al.*, 2001).

Consistent with its key role in fundamental cellular processes and the high degree of subunit structural and functional conservation in Opisthokonts, it is generally thought that the core-TFIIH is highly conserved in Eukaryotic lineages. Nevertheless, in spite of the considerable number of sequenced genomes available, no extensive *in silico* investigation has been performed on the eukaryotic kingdom. Only a few genomes of parasitic intracellular organisms have been investigated. In *P. falciparum*, a two dimensional Hydrophobic Cluster Analysis combined with profile-based searches identified the complete core-TFIIH (Callebaut *et al.*, 2005). In the *T. brucei* genome, the *in silico* investigation unambiguously revealed the presence of the XPB, XPD, P44 and P52 subunits (Lecordier *et al.*, 2007), whereas the P34, P8 and P62 have been recently isolated using tandem affinity purification experiments associated with two additional unknown proteins TPS1 and TPS2 (Lee *et al.*, 2009). A reduced core composed respectively of XPB, XPD, P44, P52 subunits in *G. lamblia* and of XPB, XPD, P44 subunits in *M. brevicollis,* has been identified in the course of genome annotations (Best *et al.*, 2004; King *et al.*, 2008), suggesting both the existence of a simplified transcriptional machinery in these eukaryotic species and specific distinct phylogenetic profiles for P62 and P34.

In the present study, we first established a reference multiple alignment for each of the 7 core-TFIIH protein families, including sequences from 63 organisms representing major eukaryotic phyla. The reference alignments allowed us to reliably estimate the sequence conservation of the core-TFIIH in Eukarya and to define 30 new evolutionary conserved Sequence Signature Motifs (SSMs) for each subunit.

These SSMs, together with previously identified motifs, allowed us to perform exhaustive sequence searches at both the protein and genome levels, in order to establish a reliable phylogenetic distribution of the 7 subunits and their domains in 63 genomes. This work revealed that XPB, XPD and P44 are present throughout the Eukaryotes. In contrast, P8 and P62 are absent or lack one domain in a few unicellular species dispersed throughout the eukaryotic lineage, while P52 is only absent in the species *G. lamblia* and p34 could not be detected in Trypanosomatids. We exploited the distinct phylogenetic distributions of the P8, P34, P52 and P62 subunits to gain insights into their functional roles through a subtractive comparative genomics approach. This type of *in silico* comparative analysis, also called differential genome display, is widely used to investigate prokaryotic genomes (for a recent review, see Barh et al, Drug Development Research, 2011) and has also been validated in Eucarya (see for instance Li et al., Cell 2004). In our study, the subtractive approach confirms the involvement of the P52 subunit in DNA repair process and suggests that the poorly documented P34 subunit is linked to mRNA processing through functional interactions with splicing factors.

2. Results

2.1. Family analysis of the seven core-TFIIH subunits

We studied the sequence conservation of the seven families of the core-TFIIH subunit in Eukaryotes, by retrieving and analysing the protein sequences from 63 species (supplementary dataset S1) representative of the main eukaryotic super-groups (Adl *et al.*, 2005), namely the Opisthokonta, the Archaeplastida and 15 protists including 2 Amoebozoa, 4 Excavata and 8 Chromalveolata. The sequences detected by Blastp searches were used to build a Multiple Alignment of Complete Sequences (MACS) for each subunit (MACS are available online at http://lbgi.igbmc.fr/puzz/index.php). Manual examination of the MACS indicated that 41 predicted protein sequences appeared to be incomplete and/or contained improperly assigned portions. For example, the comparison of transcript and protein sequences from *C. elegans* and *C. briggsae* revealed that the XPD predicted protein of *C. elegans* exhibited numerous insertions/deletions, resulting from erroneous intron/exon predictions. Another example is the P62 sequence of *M. musculus,* which lacked the N-terminal region. Manual examination of genomic and transcript sequences showed that the protein sequence could in fact be extended by 26 residues, suggesting a gene/protein prediction error.

Orthologs were determined by defining Short Signature Motifs (SSMs) for each subunit that encompass known but also newly characterized conserved motifs distributed throughout the primary sequence (figure 1 and supplementary dataset S2). We identified 9 and 2 SSMs in the P52 and P8 proteins respectively, for which only short interaction regions had been structurally characterized (Vitorino *et al.*, 2007), 8 new SSMs in P34 for which a single C4 zinc finger motif located at the C terminus has been previously identified and 3 additional SSMs for P62, including a motif similar to the BSD domain (Doerks *et al.*, 2002) that we called the BSD-like motif. We also defined 4, 8 and 5 new SSMs for the best characterized subunits, XPB, XPD and P44 respectively.

In addition, extensive BLAST searches at the genomic level, using selected sequence portions encompassing one or several SSMs, were required in order to define both the exact sequence and the complete set of the P62, P52, P34 and P8 subunits. Accession numbers of proteins or genomic locations are provided in supplementary dataset S3. This in-depth investigation allowed us to identify 12 P8 genes, which were not previously predicted probably because of the small size of the coding sequence. It also allowed us to establish the absence of domains or subunits in some species (see below) at both the protein and genomic levels.

Sequence conservation analyses showed that the XPB, XPD and P44 catalytic subunits are the most conserved subunits within the core-TFIIH with 50%, 52% and 35% mean residue identity respectively. In contrast, the P62 family shows only 19% mean residue identity for the selected set of species, revealing a surprising variability, even compared to the other non-enzymatic subunits (30%, 30% and 27% for P34, P8 and P52 respectively).

2.2. Phylogenetic distribution of the core-TFIIH subunits

The phylogenetic distribution shown in figure 2 revealed that the core-TFIIH is highly conserved among Eucarya. The catalytic subunits, the XPD and XPB helicases and the ubiquitin ligase P44, are present in all studied species. P52 is missing in a single species, *G. lamblia*. P34 appears to be conserved in all investigated species, except the Euglenozoa. In fact, sequence analysis of the potential p34 proteins identified by tandem affinity purification experiments in *T. brucei* (Tb11.01.7730), *T. cruzei* (Tc 00.104705350870.14) and *L. major* (Lmj F32.2885) (Lee *et al.*, 2009) revealed several insertions/deletions, notably in the canonical C4 zinc-finger motif and the absence of most of the SSMs. Thus, these genes constitute either a non-orthologous displacement or have diverged beyond recognition. In both cases, they reflect the presence of an atypical P34 in the core-TFIIH of trypanosomatids. Interestingly, P62 is absent in three unicellular organisms, the two amitochondriate organisms (*G. lamblia* and *E. cuniculi*), and the choanoflagellate *M. brevicolli*s, a free living Opisthokont. In addition, the ortholog found in *E. histolytica* clearly lacks the N-terminus PH/PTB domain, suggesting a partial loss of function for P62 in this particular organism. Surprisingly, the non essential P8 protein is only absent in *M. brevicollis* and *E. cuniculi*.

2.3. Subtractive analysis

In view of the absence of the non-catalytic core-TFIIH subunits in some organisms, we used a comparative genomic approach, based on proteome subtraction to investigate potential additional roles for these proteins. The basic assumption of the subtractive approach is that proteins that function together in a pathway or structural complex tend to co-evolve, i.e. to be present in the same set of species (Pellegrini *et al.*, 1999). The approach involves identifying proteins that exhibit a presence/absence pattern similar to the target protein in a subset of species. To perform our analysis, we chose phylogenetically distant organisms that have well documentated proteomes of similar size. As a reference set, we considered proteins conserved between Opisthokonta and Chromalveolata, i.e. the *S. cerevisiae* proteins conserved in *T. parva*. Comparisons with additional organisms were then performed to delineate *S. cerevisiae* gene sets exhibiting presence/absence profiles similar to P8, P62 and P52 (figure 2). Finally, to gain insight into the putative function of P34, we hypothesized that the remarkable sequence divergence of the potential Trypanosomatid counterparts was likely to indicate the absence of this subunit in this taxon and therefore, we searched for genes conserved in *S. cerevisiae* and *T. parva* but absent in *T. brucei*.

2.4. Identification of the co-evolving proteins of P8, P62 and P52

To delineate the respective co-evolving protein sets of P8, P62 and P52, we compared our reference set, i.e. *S. cerevisiae* proteins conserved in *T. parva*, with three additional proteomes exhibiting differential gene losses for the considered subunits: *G. lamblia*, *E. cuniculi* and *M. brevicollis*. These comparisons resulted in the detection of 36, 102 and 137 genes with a presence/absence pattern similar to the P8, P62 and P52 subunits, respectively (supplementary dataset S4).

The GO annotations for the Biological Process ontology (BP5) revealed an enrichment in genes involved in :

i) biosynthetic processes related to ribonucleotide (GO:0009260; P-value=2,23.10^{-9}), and D-ribose (GO:0019302; P-value=3,32.10^{-6}) and phospholipid transport (GO:0015914; P-value=1,16.10^{-4}) for the 36 genes sharing the P8 distribution (i.e. present in *S. cerevisiae*, *T. parva*, *G. lamblia* and absent in *E. cuniculi* and *M. brevicollis*) (supplementary dataset S5);

ii) coenzyme catabolism and energetic metabolism processes (GO:0009109; P-value=2,61.10^{-7} and GO:0045333; P-value=6,71.10^{-7}) for the 102 genes sharing the P62 distribution (i.e. present in *S. cerevisiae*, *T. parva* and absent in *G. lamblia*, *E. cuniculi* and *M. brevicollis*) (supplementary dataset S6);

iii) DNA metabolism (GO:0006308; P-value=2,1.10^{-5}; GO:0009263 P-value=2,1.10^{-5}), DNA repair (GO:0006281; P-value=2,2.10^{-4}) and transcription DNA dependent processes (GO:0006351; P-value=7,6.10^{-4}) for the 137 genes sharing the P52 distribution (supplementary dataset S7). Among these 137 genes, 19 are involved in DNA repair and 17 in transcription DNA dependent processes (see short descriptions in supplementary data S8). DNA repair concerns the maintenance of genomic integrity and includes distinct repair pathways corresponding to specific DNA damage: the NER, the Base Excision Repair (BER) and the Double Strand Break DNA repair (DSBR). In this context, we note that 11 of the 19 DNA repair genes are involved in DSBR and/or NER pathways (MEC1, TEL1, RAD50, MRE11, SMC5, SMC6 and MSH3, RAD1, RAD2, TFB3, RFA1 respectively), while 9 genes are more specifically linked to the RNA polII transcriptional pathway (SPT5, RBP7, TBP, TFB3, CCR4, RAD2, DST1, TF2B, ESS1) and 6 participate in chromatin remodelling and histone modifications during RNA polII transcription or DNA repair (SET2, SPT16, Pob3, ASF1, MEC1, TEL1).

2.5. Identification of the co-evolving proteins of P34

A set of 260 genes with a presence/absence pattern similar to P34 (i.e. present in *S. cerevisiae* and *T. parva* but absent in *T. brucei*) were detected (figure 3A). GO annotation of this gene list for the Biological Process ontology (BP5) indicates that 107 genes are linked to the RNA metabolic process, with a highly significant enrichment in pathways related to the RNA process (P-values<10^{-13}), mRNA metabolic process (GO:0016071; P-value=1.85.10^{-16}), RNA splicing (GO:0008380; P-value=1.54.10^{-16}) and mRNA processing (GO: 0006397; P-value=1.48.10^{-19}) (figure 3B and Supplementary dataset S9). Among these 107 genes, 19 genes participate in rRNA or tRNA processing, 41 genes in mRNA processing and 30 in RNA PolII mediated transcription and/or its regulation, including CCL1 and TFB3, two subunits of the CAK subcomplex (for more details, see supplementary dataset S10*)*.

Among the 41 genes involved in mRNA processing, 2 genes participate in capping/decapping (CEG1, DCP2) and 6 in polyadenylation (NAB2, RNA14, PTA1, PAN2, CFT2, TIF4631), while 33 participate in intron splicing. Table 2 describes some well documented splicing genes that include compounds of the U1, U2 and U4/U5/U6 snRNP complexes, as well as major proteins transiently associated with the spliceosome that participate in the remodelling of spliceosome content during the splicing cycle**.** It is worth noting that 7 genes (LUC7, RU1C, PRP40, BBP, PRP16, SLU7, PRP28) participate in the recognition of the 5’ or 3’ single strand of the intron in the earliest step of the splicing cycle (Wahl *et al.*, 2009).

3. Discussion

3.1. New motifs in the core-TFIIH: reappraisal of the subcomplex evolution

In this study, we have defined 39 new Short Signature Motifs that characterize the 7 protein families of the core-TFIIH. Together with the 29 previously known motifs, they now allow a precise delineation of protein families. These new motifs are particularly beneficial for the poorly characterized P52, P34 and P8 sequence families. The analysis at the protein sequence level was completed by genomic searches to retrieve the full complement of sequence orthologs, including missed and badly predicted genes. The results of this combined approach and our manual curation highlight the importance of gene prediction errors in eukaryotic genomes, which can considerably hamper knowledge extraction in comparative genomic studies. The P8 family constitutes a striking example: 19% of these genes were not predicted in the investigated species, leading to an apparently sparse and erratic distribution. We hope that the newly defined motifs and the multiple alignments of the curated sequences, which are acccessible *via* a user-friendly web site, will constitute a valuable resource for future studies of the core-TFIIH.

The manually verified phylogenetic distribution of the core-TFIIH subunits indicates that only four subunits (p8, p34, p52, p62) are missing in a few organisms. Of these, p8 was shown to be accessory in yeast while the other three, which are essential in yeast, are known to play a structural role in TFIIH complex formation. The obtained distribution reveals the high conservation of this complex among eukarya, which is consistent with its vital biological roles. These results contrast with previous studies (Best *et al.*, 2004; Callebaut *et al.*, 2005; King *et al.*, 2008; Lecordier *et al.*, 2007; Morrison *et al.*, 2007) that suggested a rudimentary basal initiation apparatus composed of a reduced core-TFIIH, especially in *G. lamblia* (XPB, XPD, P44, P34) and *M. brevicollis* (XPB, XPD, P44). In fact, only two genes are lacking in *G. lamblia* (P52 and P62) and in *E. cuniculi* (P8 and P62), despite the compact genomes of these species. Two genes (P8 and P62) are also absent in *M. brevicollis*, an Opisthokont that belongs to the closest lineage of Metazoa (King *et al.*, 2008). In this organism, the genome analysis indicates the absence of most intercellular signalling pathways, as well as of various transcription factors, co-activators and chromatin remodelling complexes, which could be consistent with the absence of P62, a subunit interacting with transcriptional activators (Kwek *et al.*, 2004). In addition, our study revealed the absence of the PH/PTB domain in the *E. histolytica* P62 ortholog, which may have functional implications for the TFIIH complex in this organism. Finally, it should be stressed that extensive divergence is observed in all the primary sequences of the potential P34 orthologs reported in Trypanosomatids.

Interestingly, we noticed a correlation between subunit distribution and sequence conservation. As expected, the three catalytic subunits (XPB, XPD and P44) are present in all investigated species and exhibit the highest sequence conservation. In contrast and somewhat surprisingly, P62 is missing in three species belonging to divergent phyla (Excavata, Fungi and Choanoflagellates), suggesting three independant gene loss events and is by far the least conserved subunit (19% identity). In comparison, p8, which acts as an accessory protein in NER (Ranish *et al.*, 2004) and is not essential for cell viability, exhibits 30% identity.

3.2. Known and potential roles of P52 in transcription and repair

The subtractive analysis, performed with the proteomes of *S. cerevisiae*, *T. parva*, *G. lamblia*, *E. cuniculi* and *M. brevicollis* to detect genes exhibiting the same phylogenetic distribution as P8, P62 and P52, provides contrasting results. The 36 and 102 genes sharing the same pattern as P8 and P62 respectively, show no statistically significant enrichment in functions potentially linked to TFIIH, although some individual genes were found that were related to transcriptional processes.

In contrast, the functional annotation of the 137 genes coevolving with P52 reveals a significant enrichment in genes involved in DNA repair or transcription processes (Supplementary dataset S8). DNA repair involves the recognition of DNA lesions, through a specific lesion sensor that in turn activates specific DNA repair mechanisms, such as NER or DSBR, as well as additional protection pathways, such as chromatin remodelling, apoptosis or transcription. The efficiency of DNA repair largely depends on the chromatin architecture that facilitates the access of the repair machinery to the DNA lesions (Altaf *et al.*, 2007; Faucher & Wellinger, 2010; Osley & Shen, 2006). Our comparative genomic approach identified 19 genes involved in DNA repair, including 6 and 5 genes that participate in DSBR and NER respectively, and 3 genes involved in chromatin remodeling or histone modifications (supplementary dataset S8). The NER pathway involves three major steps: the formation of the pre-incision complex at the damage sites, including the entire complex TFIIH, the excision of the oligonucleotide stretch of single stranded DNA by specific endonucleases and the re-synthesis and ligation of a DNA patch to fill the gap. Of the detected genes, 5 participate in the pre-incision step (TFB2/P52, TFB3/MAT1, RAD1/XPF and RAD2/XPG, RFA1/RPA1). Among these, RAD1/XPF and RAD2/XPG catalyse the incision in the 3’ and 5’ sides of the lesion and RFA1 facilitates the recruitment of these endonucleases to the DNA damage. Interestingly, during NER, the anchoring of TFIIH to DNA requires the ATPase activity of XPB, which is regulated through a strong interaction with P52 (Coin *et al.*, 2007) and the open DNA structure generated by the TFIIH enables the recruitment of RFA1/RPA1, XPG/RAD2 and XPF/RAD1 (Fagbemi *et al.*, 2011). Thus, our computational analysis clearly confirms the reported regulatory functions of P52 in DNA repair and more precisely, in the NER pathway.

3.3. Predictives roles of P34 in splicing

The subtractive analysis also identified 260 genes that coevolved with P34 (i.e. conserved in Opisthokonts and Chromalveolates and absent in trypanosomatids). The functional annotation of this gene set indicates a significant enrichment in the splicing process (P-values=<10^{-13}). Intrigingly, some of these genes belong to the U1snRNP complex (LUC7, PRP40, RU1C, BBP) and play a major role in the selection of the 5’ single strand or the stability of the U1snRNA-5’ single strand interaction (Table 2). Taken together, these results suggest a possible role for P34 in splicing mechanisms, which, like transcription and mRNA processes, are known to be atypical in Trypanosomatids (Liang *et al.*, 2003). Indeed, a majority of individual mRNAs possess an unusual 5’ terminal capped structure and are resolved by spliced leader (SL) *trans* splicing from polycistronic pre-mRNA (Gunzl, 2010). The *cis* and *trans* splicing are carried out by a unique spliceosomal machinery characterized by: i) the full set of the five U snRNAs that are shorter and deviate from human counterparts (Liang *et al.*, 2003), ii) the essential role of U1 snRNA for *cis* splicing but not for *trans* splicing, iii) the presence of snRNP Trypanosome specific splicing factors and iv) some conventional splicing factors that evolved to carry out distinct and specific functions in Trypanosomatids, such as U1A, a compound of U1snRNP that is involved in *trans* splicing and polyadenylation but not in *cis* splicing (Gunzl, 2010; Tkacz *et al.*, 2010).

Moreover, several lines of evidence suggest that transcription and splicing are tighly coupled. The binding of 5’U1 snRNA to the 5’ promoter proximal intron may enhance the transcription level independently of the splicing events in the context of U1 snRNP (Alexander *et al.*, 2010; Furger *et al.*, 2002). Some reports also suggest that the general transcription factor THIIH could participate in tis coupling : i) XPB is increased 3 fold at the wild type 5’ splice site promoter relative to the mutated 5’ splice site promoter (Damgaard *et al.*, 2008); ii) the trypasomatid XPB counterpart is associated with the SMD3 protein, a spliceosomal core protein that binds U1 snRNA (Tkacz *et al.*, 2010); iii) the purified preparation of the entire TFIIH complex contains a stoechiometric amount of U1snRNA that specifically associates with the CYCLIN H (Kwek *et al.*, 2002); iv) the interaction between CYCLIN H and U1snRNA (O'Gorman *et al.*, 2005) enhances transcription initiation and re-initiation from the scaffold complex (Kwek *et al.*, 2002) and is mediated by the U1 snRNA Stem Loop II that is absent in *T. brucei* (Liang *et al.*, 2003), like the P34 and CYCLIN H proteins.

In this context, our *in silico* results not only suggest for the first time a functional link between the P34 subunit of TFIIH, the splicing factors and the U1 snRNA, but also allows to hypothesize that P34 might be involved either in the earlier first step of mRNA splicing or in the U1snRNA enhancement of transcription that requires the stem loop II U1 snRNA secondary structure, snRNP proteins (Alexander *et al.*, 2010) and the 5' single strand of the promoter proximal intron (Furger *et al.*, 2002). This latter hypothesis might be in agreement with recent studies showing that TAF15, a transitory partner of the general transcription factor TFIID is associated with a fraction of human U1snRNA and might regulate the level of free U1snRNA (Jobert *et al.*, 2009; Kugel & Goodrich, 2009).

4. Conclusions

In this study, we describe an exhaustive study of the phylogenetic distribution of the 7 subunits of the core-TFIIH. Our results indicate first, that the core-TFIIH is more conserved in Eucarya than previously reported with only 3 genes, namely P8, P62 and P52, lacking in a few rare species, and second, the absence of a P62 functional module in the *E. histolytica* species, and third, the presence of extremely divergent P34 proteins in Trypanosomatids.
Our subtractive analysis confirms the role of P52 in DNA repair and suggests for the first time that P34 may be involved in the earlier first step of splicing or in U1 snRNA enhancement of transcription. In agreement with the new paradigm emphasizing the large plasticity of the TFIIH complex (for more details, see recent review (Egly & Coin, 2011), this surprising finding indicates new directions for P34 related investigations. Notably, it will be of major interest to establish whether P34 is a reliable actor of the splicing and/or transcriptional enhancement processes, as well as to decipher whether its putative activity is performed only within the core-TFIIH or in other non-TFIIH complexes.

5. Materials and methods

5.1. Sequence family analysis and phylogenetic distribution

Sequences of the core-TFIIH proteins were examined in 63 eukaryotic organisms with complete genome sequences: 15 Metazoa, 26 Fungi, 7 Archaeplastida (Viridiplantae) and 15 Protists (4 Excavata, 8 Chomalveolata, 2 Amoeboza and 1 Choanoflagellida, a close lineage of Metazoa). The complete list of species is provided in the supplementary dataset S1.

Initial BlastP searches (Altschul *et al.*, 1997) were conducted at the National Center for Biotechnology Information site (http://www.ncbi.nlm.nih.gov/BLAST) in the non-redundant protein database (E =< 0.001) using *S. cerevisiae* proteins as queries: XPD (**P06839**), XPB (**Q00578**), P62 (**P32776**), P52 (**Q02939**), P44 (**Q04673**), P34 (**Q12004**) and P8 (**Q3E7C1**). When initial searches failed to recover a protein candidate, sequences from a close relative of the target genome were used to identify the counterpart using TBlastN from the NCBI site or the dedicated websites given in supplementary dataset S11. BLAST parameters (Expect threshold and filtering options) were adapted if needed for short and/or biased sequences.

For each subunit, the likely homologous sequences detected by BLAST searches were aligned using PipeAlign (Plewniak *et al.*, 2003). Based on secondary structure and known Sequence Signature Motifs (SSMs), each alignment was manually refined and false-positive protein sequences were removed. The complete alignments of the core-TFIIH subunits are available at http://bips.u-strasbg.fr/coreTFIIHalignment. From each alignment, we defined new SSMs that include at leat 4 conserved amino acid residues or exhibit similar physico-chemical properties in 90% of aligned sequences. The SSM sequences are specified in supplementary dataset S2. Sequence conservation within each family was estimated by calculating the pairwise sequence identities between complete sequences from 44 organisms with the full set of core-TFIIH subunits (13 metazoans, 6 plants, 13 fungi and 9 protists; see supplementary dataset S12).

5.2. Subtractive analysis.

Subtractive analyses were performed using the Orthoinspector software suite (Linard *et al.*, 2011) that detects orthology and inparalogy relationships between species by analysing BLAST all-against-all searches. Sets of genes with suitable phylogenetic profiles were then analysed using the integrated gene annotation database, DAVID 6.7 (the Database for Annotation, Vizualization and Integration Discovery) (Sherman *et al.*, 2007), which provides a Gene Ontology (GO) term enrichment analysis tool. Only GO term enrichments with P-values $<10^{-3}$ were considered.

Acknowledgements

We would like to thank the members of the Strasbourg Bioinformatics Platform (BIPS) for their support. We acknowledge financial support from the ANR (Puzzle-Fit: 09-PIRI-0018-02) and Institute funds from the CNRS, INSERM, and the Université de Strasbourg.

References

Adl S. M., Simpson A. G., Farmer M. A., Andersen R. A., Anderson O. R., Barta J. R., Bowser S. S., Brugerolle G., Fensome R. A., Fredericq S., James T. Y., Karpov S., Kugrens P., Krug J., Lane C. E., Lewis L. A., Lodge J., Lynn D. H., Mann D. G., McCourt R. M., Mendoza L., Moestrup O., Mozley-Standridge S. E., Nerad T. A., Shearer C. A., Smirnov A. V., Spiegel F. W., and Taylor M. F. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* **52:** 399-451.

Aguilar-Fuentes J., Valadez-Graham V., Reynaud E., and Zurita M. (2006). TFIIH trafficking and its nuclear assembly during early Drosophila embryo development. *J Cell Sci* **119:** 3866-75.

Alexander M. R., Wheatley A. K., Center R. J., and Purcell D. F. (2010). Efficient transcription through an intron requires the binding of an Sm-type U1 snRNP with intact stem loop II to the splice donor. *Nucleic Acids Res*.

Altaf M., Saksouk N., and Cote J. (2007). Histone modifications in response to DNA damage. *Mutat Res* **618:** 81-90.

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., and Lipman D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Best A. A., Morrison H. G., McArthur A. G., Sogin M. L., and Olsen G. J. (2004). Evolution of eukaryotic transcription: insights from the genome of Giardia lamblia. *Genome Res* **14:** 1537-47.

Callebaut I., Prat K., Meurice E., Mornon J. P., and Tomavo S. (2005). Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics* **6:** 100.

Coin F., Oksenych V., and Egly J. M. (2007). Distinct roles for the XPB/p52 and XPD/p44 subcomplexes of TFIIH in damaged DNA opening during nucleotide excision repair. *Mol Cell* **26:** 245-56.

Coin F., Oksenych V., Mocquet V., Groh S., Blattner C., and Egly J. M. (2008). Nucleotide excision repair driven by the dissociation of CAK from TFIIH. *Mol Cell* **31:** 9-20.

Damgaard C. K., Kahns S., Lykke-Andersen S., Nielsen A. L., Jensen T. H., and Kjems J. (2008). A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol Cell* **29:** 271-8.

Doerks T., Huber S., Buchner E., and Bork P. (2002). BSD: a novel domain in transcription factors and synapse-associated proteins. *Trends Biochem Sci* **27:** 168-70.

Egly J. M., and Coin F. (2011). A history of TFIIH: two decades of molecular biology on a pivotal transcription/repair factor. *DNA Repair (Amst)* **10:** 714-21.

Fagbemi A. F., Orelli B., and Scharer O. D. (2011). Regulation of endonuclease activity in human nucleotide excision repair. *DNA Repair (Amst)* **10:** 722-9.

Faucher D., and Wellinger R. J. (2010). Methylated H3K4, a transcription-associated histone modification, is involved in the DNA damage response pathway. *PLoS Genet* **6**.

Fribourg S., Romier C., Werten S., Gangloff Y. G., Poterszman A., and Moras D. (2001). Dissecting the interaction network of multiprotein complexes by pairwise coexpression of subunits in E. coli. *J Mol Biol* **306:** 363-73.

Furger A., O'Sullivan J. M., Binnie A., Lee B. A., and Proudfoot N. J. (2002). Promoter proximal splice sites enhance transcription. *Genes Dev* **16:** 2792-9.

Gervais V., Lamour V., Jawhari A., Frindel F., Wasielewski E., Dubaele S., Egly J. M., Thierry J. C., Kieffer B., and Poterszman A. (2004). TFIIH contains a PH domain involved in DNA nucleotide excision repair. *Nat Struct Mol Biol* **11:** 616-22.

Gunzl A. (2010). The pre-mRNA splicing machinery of trypanosomes: complex or simplified? *Eukaryot Cell* **9:** 1159-70.

Hong S. W., Hong S. M., Yoo J. W., Lee Y. C., Kim S., Lis J. T., and Lee D. K. (2009). Phosphorylation of the RNA polymerase II C-terminal domain by TFIIH kinase is not essential for transcription of Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A* **106:** 14276-80.

Jawhari A., Boussert S., Lamour V., Atkinson R. A., Kieffer B., Poch O., Potier N., van Dorsselaer A., Moras D., and Poterszman A. (2004). Domain architecture of the p62 subunit from the human transcription/repair factor TFIIH deduced by limited proteolysis and mass spectrometry analysis. *Biochemistry* **43:** 14420-30.

Jobert L., Pinzon N., Van Herreweghe E., Jady B. E., Guialis A., Kiss T., and Tora L. (2009). Human U1 snRNA forms a new chromatin-associated snRNP with TAF15. *EMBO Rep* **10:** 494-500.

Kanin E. I., Kipp R. T., Kung C., Slattery M., Viale A., Hahn S., Shokat K. M., and Ansari A. Z. (2007). Chemical inhibition of the TFIIH-associated kinase Cdk7/Kin28 does not impair global mRNA synthesis. *Proc Natl Acad Sci U S A* **104:** 5812-7.

King N., Westbrook M. J., Young S. L., Kuo A., Abedin M., Chapman J., Fairclough S., Hellsten U., Isogai Y., Letunic I., Marr M., Pincus D., Putnam N., Rokas A., Wright K. J., Zuzow R., Dirks W., Good M., Goodstein D., Lemons D., Li W., Lyons J. B., Morris A., Nichols S., Richter D. J., Salamov A., Sequencing J. G., Bork P., Lim W. A., Manning G., Miller W. T., McGinnis W., Shapiro H., Tjian R., Grigoriev I. V., and Rokhsar D. (2008). The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451:** 783-8.

Kugel J. F., and Goodrich J. A. (2009). In new company: U1 snRNA associates with TAF15. *EMBO Rep* **10:** 454-6.

Kwek K. Y., Murphy S., Furger A., Thomas B., O'Gorman W., Kimura H., Proudfoot N. J., and Akoulitchev A. (2002). U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat Struct Biol* **9:** 800-5.

Kwek K. Y., O'Gorman W., and Akoulitchev A. (2004). Transcription meets DNA repair at a PH domain. *Nat Struct Mol Biol* **11:** 588-9.

Lecordier L., Devaux S., Uzureau P., Dierick J. F., Walgraffe D., Poelvoorde P., Pays E., and Vanhamme L. (2007). Characterization of a TFIIH homologue from Trypanosoma brucei. *Mol Microbiol* **64:** 1164-81.

Lee J. H., Jung H. S., and Gunzl A. (2009). Transcriptionally active TFIIH of the early-diverged eukaryote Trypanosoma brucei harbors two novel core subunits but not a cyclin-activating kinase complex. *Nucleic Acids Res* **37:** 3811-20.

Liang X. H., Haritan A., Uliel S., and Michaeli S. (2003). trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell* **2:** 830-40.

Linard B., Thompson J. D., Poch O., and Lecompte O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. *In* "BMC Bioinformatics", pp. 11.

Matsui P., DePaulo J., and Buratowski S. (1995). An interaction between the Tfb1 and Ssl1 subunits of yeast TFIIH correlates with DNA repair activity. *Nucleic Acids Res* **23:** 767-72.

Morrison H. G., McArthur A. G., Gillin F. D., Aley S. B., Adam R. D., Olsen G. J., Best A. A., Cande W. Z., Chen F., Cipriano M. J., Davids B. J., Dawson S. C., Elmendorf H. G., Hehl A. B., Holder M. E., Huse S. M., Kim U. U., Lasek-Nesselquist E., Manning

G., Nigam A., Nixon J. E., Palm D., Passamaneck N. E., Prabhu A., Reich C. I., Reiner D. S., Samuelson J., Svard S. G., and Sogin M. L. (2007). Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. *Science* **317:** 1921-6.

O'Gorman W., Thomas B., Kwek K. Y., Furger A., and Akoulitchev A. (2005). Analysis of U1 small nuclear RNA interaction with cyclin H. *J Biol Chem* **280:** 36920-5.

Osley M. A., and Shen X. (2006). Altering nucleosomes during DNA double-strand break repair in yeast. *Trends Genet* **22:** 671-7.

Pellegrini M., Marcotte E. M., Thompson M. J., Eisenberg D., and Yeates T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-8.

Plewniak F., Bianchetti L., Brelivet Y., Carles A., Chalmel F., Lecompte O., Mochel T., Moulinier L., Muller A., Muller J., Prigent V., Ripp R., Thierry J. C., Thompson J. D., Wicker N., and Poch O. (2003). PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res* **31:** 3829-32.

Ranish J. A., Hahn S., Lu Y., Yi E. C., Li X. J., Eng J., and Aebersold R. (2004). Identification of TFB5, a new component of general transcription and DNA repair factor IIH. *Nat Genet* **36:** 707-13.

Rouillon C., and White M. F. (2011). The evolution and mechanisms of nucleotide excision repair proteins. *Res Microbiol* **162:** 19-26.

Schroeder S. C., Schwer B., Shuman S., and Bentley D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev* **14:** 2435-40.

Sherman B. T., Huang da W., Tan Q., Guo Y., Bour S., Liu D., Stephens R., Baseler M. W., Lane H. C., and Lempicki R. A. (2007). DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8:** 426.

Tkacz I. D., Gupta S. K., Volkov V., Romano M., Haham T., Tulinski P., Lebenthal I., and Michaeli S. (2010). Analysis of spliceosomal proteins in Trypanosomatids reveals novel functions in mRNA processing. *J Biol Chem* **285:** 27982-99.

Tremeau-Bravard A., Perez C., and Egly J. M. (2001). A role of the C-terminal part of p44 in the promoter escape activity of transcription factor IIH. *J Biol Chem* **276:** 27693-7.

Viladevall L., St Amour C. V., Rosebrock A., Schneider S., Zhang C., Allen J. J., Shokat K. M., Schwer B., Leatherwood J. K., and Fisher R. P. (2009). TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. *Mol Cell* **33:** 738-51.

Vitorino M., Coin F., Zlobinskaya O., Atkinson R. A., Moras D., Egly J. M., Poterszman A., and Kieffer B. (2007). Solution structure and self-association properties of the p8 TFIIH subunit responsible for trichothiodystrophy. *J Mol Biol* **368:** 473-80.

Wahl M. C., Will C. L., and Luhrmann R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell* **136:** 701-18.

Zurita M., and Merino C. (2003). The transcriptional complexity of the TFIIH complex. *Trends Genet* **19:** 578-84.