STRUCTURE AND FUNCTION OF EUKARYOTIC TRANSCRIPTION REGULATORY COMPLEXES

Université de Strasbourg

Mémoire présenté pour l'obtention de

L'Habilitation à diriger des recherches

Spécialité Biologie Structurale par

Gabor PAPAI, PhD Chargé de recherche de première classe

soutenu à Strasbourg le 12 février 2018

Commission d'examen:

MARC TIMMERS (DKFZ, FREIBURG, GERMANY)	RAPPORTEUR EXTERNE
PATRICK BRON (CBS, MONTPELLIER, FRANCE)	RAPPORTEUR EXTERNE
DIDIER DEVYS (IGBMC, STRASBOURG, FRANCE)	RAPPORTEUR INTERNE
CATHERINE VENIEN-BRYAN (UPMC, FRANCE)	EXAMINATEUR
JEAN CAVARELLI (IGBMC, STRASBOURG, FRANCE)	EXAMINATEUR
PATRICK SCHULZ (IGBMC, STRASBOURG, FRANCE)	GARANT D'HABILITATION

TABLE OF CONTENTS

Acknowledgement 3

Introduction of the candidate – summary of past and proposed work 4

Curriculum vitae 5

Past research activities 8

Introduction **8**

Transcription initiation 8

Chromatin 11

Results **12**

Mapping the initiator binding Taf2 subunit in the structure of hydrated yeast TFIID 12

TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation 25

The architecture of human general transcription factor TFIID core complex 32

Molecular structure of promoter-bound yeast TFIID 39

Structure and dynamics of a 197 base-pair nucleosome in complex with linker histone H1 52

Future project 66

Structure and function of endogenous TFIID-containing transcription preinitiation complex **67**

References 76

ACKNOWLEDGEMENT

All the studies I made were carried out in teamwork and all the team members' contributions were important. Therefore, I would like to thank everybody, especially the members of Boros and Schultz lab, who helped me to conduct my research.

I am grateful to my garant, Dr Patrick Schultz, for accepting me in his team and for teaching me how to solve structures using an electron microscope. I am also thankful for his constant support.

I would like to thank to my jury, Dr Catherine Venien-Bryan, Dr Marc Timmers, Dr Patrick Bron, Dr Dider Devys and Dr Jean Cavarelli for participating in my HDR defence and to accept to review the hereby presented work.

I am grateful to my thesis supervisor, Dr. Imre Boros, who introduced me to practical research, who followed and supported my early, not-always successful trials at the lab bench. I greatly appreciate his mentorship.

I am thankful to Dr Andor Udvardy, who taught me to be patient and thorough with my experiments. It was an immense pleasure to work with him.

Finally, I'm especially thankful for the constant support of my family: my wife Judit, and our two daughters, Zsofia and Dora.

INTRODUCTION OF THE CANDIDATE – SUMMARY OF PAST AND PROPOSED WORK

I obtained my master degree from the József Attila University (Szeged, Hungary) in molecular biology and biotechnology. During my cursus, I became enthralled with the complexity and elegancy of the eukaryotic transcription machinery. As a master student, I joined the laboratory of Dr. Imre Boros to study transcription regulation using molecular biology methods. I continued in the same laboratory as a graduate student to study the intriguing aspects of how the cell orchestrates gene transcription in *Drosophila melanogaster*. I participated in the finding of the higher-eukaryotic homologues of the yeast transcriptional co-activator SAGA and we showed that it's acetylase function was divided into two homologue complexes, SAGA and ATAC. I studied the promoter regulation of the gene encoding Ada2a, a subunit of the ATAC complex.

During my thesis work I learned a lot of techniques and I became more and more convinced that the visualization of the transcriptional machines will lead me closer to the understanding of their function. After obtaining my PhD diploma I joined the laboratory of Dr. Patrick Schultz where I could learn structural biology to be able to visualize protein complexes. I have been exploring the structure and function of the general transcription factor TFIID since then. To obtain macromolecular structures, I'm using the single particle cryo-electron microscopy method.

I solved structures of TFIID alone or in complex with other partners, like transcriptional activator or DNA. These structures shed light on the early events of transcription initiation in yeast. The structures of different assembly stages of the human TFIID complex showed how symmetry is broken during maturation.

I continue to explore how TFIID participates in transcription initiation. TFIID is a fascinating protein complex. Its component, TBP, is involved to direct the RNA polymerase II to the transcription start site. TBP can fulfill *in vitro* this function alone, but *in vivo* it is surrounded by more than ten other proteins. Majority of these proteins are also essential, raising the question of their exact role in the cell.

Transcription is a delicately regulated process, which involves a plethora of actors, like activator proteins, coactivators, remodelers and the nucleosomes themselves. In our research team, we are interested to discover the mechanisms how the chromatin is acetylated to be prepared for transcription by SAGA or NuA4 and then how they can be repositioned on the DNA by the Swi/Snf complex.

In the following pages, I will review my past and current research activities. In the first part I will present my past research after graduation. In the second part I will describe the project I'm interested to pursue.

CURRICULUM VITAE

Personal data	Born 08/01/1974, Hungary Citizen of Hungary
Address	IGBMC Department of Integrated Structural Biology 1, rue Laurent Fries 67400 Illkirch, France Tel :+33 (0)3 69 48 52 88 e-mail: papai@igbmc.fr

Education

1994 – 1999	M.Sc. József Attila University (JATE) Faculty of Natural Sciences, Szeged
1999 – 2005	Ph.D. University of Szeged, Faculty of Natural Sciences Molecular and Cellular Biology Program, Szeged, Hungary.

Professional career

1999 – 2002	Ph.D. fellow in the Ph.D. program of University of Szeged, Biological Research Center, Szeged, Hungary
2002 – 2005	Staff scientist at the Biological Research Center, Szeged, Hungary
2005	Postdoctoral fellow at Department of Biochemistry and Molecular Biology, University of Szeged, Szeged, Hungary
2005 – 2013	Postdoctoral fellow at Department of Structural Biology and Genomics, IGBMC, Illkirch, France
2013 –	Inserm Chargé de Recherche 1ère classe (CR1) at Department of Integrated Structural Biology, IGBMC, Illkirch, France

Fellowships Awarded

1999 – 2002	Hungarian State Ph.D. Fellowship
2002 – 2005	Hungarian Academy of Sciences Young Researcher Fellowship

Publication List

Rakonczay Z. Jr, Takács T., Mándi Y., Iványi B., Varga I., Pápai G., Boros I., Lonovics J. (2001) Water immersion pretreatment decreases pro-inflammatory cytokine production in cholecystokinin-octapeptide-induced acute pancreatitis in rats: possible role of HSP72. Int J Hyperthermia. 17: 520-535

Rakonczay Z. Jr, Takács T., Iványi B., Mándi Y., Pápai G., Boros I., Varga I., Jost K., Lonovics J. (2002) The effects of hypo- and hyperthermic pretreatment on sodium taurocholate-induced acute pancreatitis in rats. Pancreas. 24: 83-89

Rakonczay Z. Jr, Takács T., Iványi B., Mándi Y., Pápai G., Boros I., Varga I., Jost K., Lonovics J. (2002) Induction of heat shock proteins fails to produce protection against trypsin-induced acute pancreatitis in rats. Clin Exp Med. 2: 89-97

Muratoglu S., Georgieva S., Pápai G., Scheer E., Enünlü I., Komonyi O., Cserpán I., Lebedeva L., Nabirochkina E., Udvardy A., Tora L., Boros I. (2003) Two different Drosophila ADA2 homologues are present in distinct GCN5 histone acetyltransferase-containing complexes. Mol Cell Biol. 23: 306-21.

Enunlu I., Papai G., Cserpan I., Udvardy A., Jeang KT., Boros I. (2003) Different isoforms of PRIP-interacting protein with methyltransferase domain/trimethylguanosine synthase localizes to the cytoplasm and nucleus. Biochem Bioph Res Co. 309: 44-51

Pápai G., Komonyi O., Tóth Z., Pankotai T., Muratoglu S., Udvardy A., Boros I. (2005) Intimate relationship between the genes of two transcriptional co-activators, ADA2a and PIMT of Drosophila. Gene. 348:13-23

Komonyi O., Pápai G., Enunlu I., Muratoglu S., Pankotai T., Kopitova D., Maróy P., Udvardy A., Boros I. (2005) DTL the Drosophila homologue of PIMT/Tgs1, nuclear receptor coactivator interacting protein/RNA methyltransferase, has essential role in development. J Biol Chem. 280:12397-12404.

Szabolcs A., Reiter R.J., Letoha T., Hegyi P., Papai G., Varga I., Jarmay K., Kaszaki J., Sari R., Rakonczay Z. Jr, Lonovics J., Takacs T. (2006) Effect of melatonin on the severity of L-arginine-induced experimental acute pancreatitis in rats. World J Gastroenterol. 12:251-8.

Letoha T., Kusz E., Pápai G., Szabolcs A., Kaszaki J., Varga I., Takács T., Penke B., and Duda E. (2006) In Vitro and in Vivo Nuclear Factor-κB Inhibitory Effects of the Cell-Penetrating Penetratin. Peptide Mol Pharmacol. 69: 2027-2036.

Crucifix C., Papai G., Schultz P. (2008) Frozen Hydrated macromolecules for structural analysis. Handbook of Cryopreparation methods for electron microscopy. CRC Press, Editors: Cavalier A, Spehner D, Humbel BM

Papai G, Tripathi MK, Ruhlmann C, Werten S, Crucifix C, Jennings JL, Link AJ, Weil PA, Schultz P. (2009) Mapping the initiator binding Taf2 subunit in the structure of hydrated yeast TFIID. Structure 17(3):363-73.

Komonyi O, Schauer T, Papai G, Deak P, Boros IM. (2009) A product of the bicistronic Drosophila melanogaster gene CG31241, which also encodes a trimethylguanosine synthase, plays a role in telomere protection. J Cell Sci. 122(Pt 6):769-74.

Cler E, Papai G, Schultz P, Davidson I. (2009) Recent advances in understanding the structure and function of general transcription factor TFIID. Cell Mol Life Sci. 66(13):2123-34.

Papai G, Tripathi MK, Ruhlmann C, Layer JH, Weil PA, Schultz P. (2010) TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation. Nature 465(7300):956-60.

Papai G, Weil PA, Schultz P. (2011) New insights into the function of transcription factor TFIID from recent structural studies. Curr Opin Genet Dev. 21(2):219-24.

Papai G, Schultz P. (2011) Transcriptional regulation by the coactivator TFIID. Med Sci (Paris). 26(12):1018-9.

Bieniossek C*, Papai G*, Schaffitzel C, Garzoni F, Chaillet M, Scheer E, Papadopoulos P, Tora L, Schultz P, Berger I. (2013) The architecture of human general transcription factor TFIID core complex. Nature, 493(7434):699-702.
*These authors contributed equally to the study

Durand A, Papai G, Schultz P. (2013) Structure, assembly and dynamics of macromolecular complexes by single particle cryo-electron microscopy. Journal of Nanobiotechnology, 11(Suppl 1):S4

Trowitzsch S, Viola C, Scheer E, Conic S, Chavant V, Fournier M, Papai G, Ebong IO, Schaffitzel C, Zou J, Haffke M, Rappsilber J, Robinson CV, Schultz P, Tora L, Berger I. (2015) Cytoplasmic TAF2-TAF8-TAF10 complex provides evidence for nuclear holo-TFIID assembly from preformed submodules. Nat Commun. 6:6011.

Pilsl M, Crucifix C, Papai G, Krupp F, Steinbauer R, Griesenbeck J, Milkereit P, Tschochner H, Schultz P. (2016) Structure of the initiation-competent RNA polymerase I and its implication for transcription. Nat Commun. 7:12126.

Roulland Y, Ouararhni K, Naidenov M, Ramos L, Shuaib M, Syed SH, Lone IN, Boopathi R, Fontaine E, Papai G, Tachiwana H, Gautier T, Skoufias D, Padmanabhan K, Bednar J, Kurumizaka H, Schultz P, Angelov D, Hamiche A, Dimitrov S. (2016) The Flexible Ends of CENP-A Nucleosome Are Required for Mitotic Fidelity. Mol Cell. 63(4):674-685.

Botte M, Zaccai NR, Nijeholt JL, Martin R, Knoops K, Papai G, Zou J, Deniaud A, Karuppasamy M, Jiang Q, Roy AS, Schulten K, Schultz P, Rappsilber J, Zaccai G, Berger I, Collinson I, Schaffitzel C. (2016) A central cavity within the holo-translocon suggests a mechanism for membrane protein insertion. Sci Rep. 6:38399.

Bednar J*, Garcia-Saez I*, Boopathi R*, Cutter AR*, Papai G*, Reymer A, Syed SH, Lone IN, Tonchev O, Crucifix C, Menoni H, Papin C, Skoufias DA, Kurumizaka H, Lavery R, Hamiche A, Hayes JJ, Schultz P, Angelov D, Petosa C, Dimitrov S. (2017) Structure and Dynamics of a 197 bp Nucleosome in Complex with Linker Histone H1. Mol Cell. 66(3):384-397.
*These authors contributed equally to the study

Sharov G, Voltz K, Durand A, Kolesnikova O, Papai G, Myasnikov AG, Dejaegere A, Ben Shem A, Schultz P. (2017) Structure of the transcription activator target Tra1 within the chromatin modifying complex SAGA. Nat Commun. 8(1):1556.

von Loeffelholz O, Papai G, Danev R, Myasnikov AG, Natchiar SK, Hazemann I, Ménétret JF, Klaholz BP. (2018) Volta phase plate data collection facilitates image processing and cryo-EM structure determination. J Struct Biol. 202(3):191-199.

Guo X, Myasnikov AG, Chen J, Crucifix C, Papai G, Takacs M, Schultz P, Weixlbaumer A. (2018) Structural Basis for NusA Stabilized Transcriptional Pausing. Mol Cell. 69(5):816-827.

Kolesnikova O, Ben ShemA, LuoJ, RanishJ, Schultz P, Papai G. (2018) Molecular structure of promoter-bound yeast TFIID. Nat Commun. doi: 10.1038/s41467-018-07096-y

Sosnowski P, Urnavicius L, Boland A, Fagiewicz R, Busselez J, Papai G, Schmidt H. (2018) The CryoEM structure of the Saccharomyces cerevisiae ribosome maturation factor Rea1. Elife. 7. pii: e39163. doi: 10.7554/eLife.39163.

PAST RESEARCH ACTIVITIES

INTRODUCTION

TRANSCRIPTION INITIATION

Eukaryotic gene expression requires the assembly of the transcription preinitiation complex (PIC) on active gene promoters. The role of the PIC, composed of ~100 proteins, is to accurately position the RNA polymerase II (Pol II) at transcription start sites (TSSs). In the last decades, PIC components have been identified and structures of the *in vitro* reconstituted core PICs have been solved. Despite intensive efforts, information on endogenous PIC composition, structure and compositional and structural dynamics is scarce. In particular no structural information is available on PICs containing the general transcription factor TFIID and on TFIID-containing PICs assembled on endogenous promoters.

Eukaryotic Pol II dependent gene transcription is a tightly regulated, essential process controlled by complex multicomponent machinery. The DNA transcribing enzyme, Pol II, cannot specifically initiate transcription at promoter sequences. *In vitro* transcription assays have shown that general transcription factors (GTFs) are required for Pol II recruitment, DNA unwinding, and accurate TSS recognition [5, 6]. One of the earliest and a highly regulated step in protein coding gene expression is the formation of a transcription preinitiation complex on the promoter DNA. The PIC consists of Pol II together with the GTFs, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH and has an approximate size of 3-4 MDa [7]. GTFs also participate in transducing the signal from cell specific transcriptional activators to Pol II and/or in Pol II promoter escape [8]. Recently other proteins/complexes were shown to be involved in transcription initiation. It was demonstrated that TFIIS is also a component of the PIC [9], while the co-activator complex SAGA is generally required for Pol II dependent transcription in yeast [10]. The Mediator complex was also reported to be present on all active gene promoters [11]. These findings suggest that *in vivo* regulated transcription initiation requires more complex machineries than originally anticipated. The exact role of these factors and the sequence of events are still poorly understood.

Nearly all biochemical and structural studies have examined transcription initiation only at TATA-containing promoters. Although TATA promoters account for only 10–30% of eukaryotic promoters [12-14], they are usually the most tightly regulated. On the other hand, TATA-less promoters often direct transcription of housekeeping genes and have heterogeneous TSSs. The biochemical studies analyzing *in vitro* transcription initiation at TATA-less promoters have been hampered by difficulties in efficient PIC assembly and low initiation activity.

Early studies using DNA–protein crosslinking of human PICs showed that Pol II is at the center of the PIC, covering over 60 bp of promoter DNA [15]. TBP, TFIIB, and TFIIF primarily contact DNA upstream of the TSS while TFIIE overlaps the TSS. TFIIH, required for promoter opening, is located at the downstream end of the PIC. Over the past 10 years, a combination of biochemical and structural studies, X-ray crystallography [16-18], and cryo-

EM [often combined with protein crosslinking coupled with mass-spectrometry (XL-MS)] of the PIC, and its intermediates in the assembly pathway [19-23], have resulted in tremendous advances in understanding the architecture of the PIC and the role of the GTFs [1, 9, 24-30].

Despite the large number of biochemical and structural studies the exact mechanistic role of each PIC component in native activated transcription initiation is still poorly understood. The first GTF to bind gene promoters is TFIID [30] that is composed of the TATA binding protein (TBP) and 13 (14 in yeast) TBP-associated factors (TAFs). Tafs are highly conserved from baker's yeast [31] to higher eukaryotes [32] and nearly all of them are essential for viability [33]. TFIID triggers PIC formation and functions also as a coactivator by interacting with transcriptional activators [34]. *In vivo* the chromatin environment modulates these interactions both negatively by nucleosome-limited access to the promoter and positively through direct interactions between transcriptional coactivator complexes, including TFIID, and modified histone tails [35].

Almost all studies analyzing the structure of the PIC used only TBP, instead of TFIID, possibly due to its large size, flexibility and fragility. In an *in vitro* system TBP alone can direct the formation of the PIC and contribute to transcription by Pol II [36], however such a system fails to respond to sequence specific trans-activators and poorly initiates transcription from promoters lacking a TATA box [37]. At the time of their discovery Tafs were described as transcriptional coactivators conveying the functional link between TBP and GTFs on one side, and transcriptional activators on the other side [38]. Biochemical and genetic information showed in different organisms that several Tafs (Taf4, Taf12 or Taf5) participate in the crosstalk between TFIID and activators such as Sp1 [39], CREB [40], p53 [41] in humans or Rap1 in yeast [42]. These observations point out the importance of Tafs in the activator dependent transcription initiation. Attempts were made to define how TFIID is involved in the PIC formation through studies deciphering TFIID interactions with activators, the promoter DNA and/or TFIIA [25, 43].

Beside their involvement in activator binding some Tafs were shown to assist TBP in promoter recognition [44, 45]. A subcomplex of human Taf1-Taf2 and TBP, but not TBP alone, was able to direct promoter selectivity of Pol II [46] underscoring the *in vivo* importance of these Tafs in promoter recognition. In this respect, some Tafs have been shown to interact with core promoter elements such as the initiator element for the C-terminus of Taf2, Taf6 with the downstream promoter element (DPE) [47] or the TATA-box for Taf4 [48] suggesting that Tafs are crucial for the binding of TFIID to promoters. Furthermore, several Tafs were shown to regulate the activity of TBP and the assembly of TFIID [29, 49-51]. In this respect TFIID harbors an auto-inhibition function that prevents TBP from binding to any TATA box in a non-regulated manner. The N-terminal domain of Taf1 is able to bind to TBP and thereby prevent its association with promoter DNA [52]. In addition, a recent study has shown that Taf11/Taf13 can bind tightly to the concave DNA-binding surface of TBP and thus displace TATA-box containing DNA from a TBP/DNA complex [29]. These results suggest that, in TFIID, several distinct TBP/Taf interactions exist, which are formed to prevent undesirable TFIID/DNA interactions, which could otherwise lead for instance to cryptic transcription initiation on genomic regions that do not contain promoter elements. The Taf1- and the Taf11/Taf13-dependent inhibition of TBP/DNA interactions suggest a further point of transcriptional control, possibly depending on promoter context or additional gene regulatory factors. TBP also interacts with the general transcription factor TFIIA, which stabilizes the TFIID–promoter complex. TFIIA and Taf1 have overlapping binding sites, thus they compete with each other for binding to TBP thereby suggesting a regulatory role of TFIIA in the binding of TFIID to the promoters [53, 54]. TFIIA was also shown to mediate the binding between transcriptional activator Rap1 and TFIID in an early transcription initiation event on the ribosomal protein encoding gene promoters [25].

A striking feature of TFIID's structural organization is that nine Tafs contain a stretch of amino acids with sequence homology to histones, including the histone fold (HF) domain involved in histone dimerization [55]. These homologies were confirmed by X-ray diffraction studies which revealed that drosophila Taf9 and Taf6 form a heterotetramer and interact through a characteristic histone fold [56] and that human Taf11 and Taf13 [57], as well as Taf4 and Taf12 [58] also contain a HF used to form heterodimers. Sequence alignment, specific heterodimerization of bacterial coexpressed Tafs and two-hybrid assays, showed that also Taf3-10 and Taf8-10 can form specific heterodimers [59, 60].

A very unique subunit stoichiometry prevails in TFIID since a subset of six subunits occurs in two copies (Taf5, 6, 9, 4, 12, 10), while the remaining seven Tafs are present as single copies. A TFIID core containing these Tafs, with the exception of Taf10, has been produced in insect cells and its cryo-EM structure showed a clear two-fold symmetry which was broken by the addition of the Taf8-10 heterodimer [26].

Yeast TFIID contains an additional subunit, Taf14 [4], which is also a constituent of six other transcription related complexes thus making its role as a bona-fide TFIID subunit difficult to evaluate. Although Taf14 is not conserved as a TFIID subunit in metazoans, its function may be retained through the chromatin binding YEATS domains that may substitute for some of the missing metazoan chromatin interaction domains such as the human TAF1 double bromodomain and the TAF3 plant homeobox domain (PHD) [61, 62]. Genetic interactions of Taf14 were described and biochemically mapped within the C-terminus of Taf2 [63].

Despite numerous efforts, TFIID is still poorly described at the structural level and no atomic model of the full complex is currently available. Inherent flexibility, poor complex stability, and sub-stoichiometric subunit composition prevented reaching high resolution structural information. Cryogenic electron microsopy (cryo-EM) and footprinting studies indicated that human TFIID can undergo massive structural rearrangements [64]. TFIID was shown to co-exist in two distinct structural states while the presence of both TFIIA and promoter DNA stabilizes a rearranged state of TFIID that enables promoter recognition and binding. A recent breakthrough was achieved by stabilizing human TFIID through its binding to TFIIA and a chimeric super core promoter (SCP) designed by combining several core promoter binding motifs found in metazoans [43]. This structure enabled the fitting of the atomic coordinates of TAF2, TAF1 and two copies of the TAF6 HEAT repeats. It also showed the interaction of TAF2 with downstream DNA promoter elements, and revealed DNA-bound TBP in a remote position where it interacts only with TFIIA.

CHROMATIN

The basic repeat unit of eukaryotic chromatin, the nucleosome, comprises a nucleosome core particle (NCP), linker DNA and a linker histone (H1) [65]. Each NCP consists of 147 DNA base pairs (bp) wrapped around a histone core octamer [66] and is separated from neighbouring NCPs by a variable length of linker DNA. Linker histones induce the formation of an apposed linker DNA stem motif [67, 68] and extend the amount of DNA protected from micrococcal nuclease digestion by ~20 bp beyond the 147 bp protected by the core octamer. Linker histones are an important determinant of nucleosome repeat length [69] and are critical for the assembly and maintenance of the 30-nm chromatin fibre [70-72] and of higher-order chromatin structures [73]. Multiple isoforms with distinct species, tissue and developmental specificity have been identified, including eleven mammalian H1 sub-types and the avian erythrocyte variant H5 [74]. Histone H1/H5 family members share a tripartite structure consisting of a conserved globular domain of ~75 residues, an N-terminal tail of 20-35 residues and a highly basic C-terminal domain (CTD) of ~100 residues [75]. Whereas the N-terminal tail has little effect on chromatin binding [75-77], the globular domain is sufficient for structure-specific nucleosome recognition and for protecting additional DNA from nuclease digestion [75, 78]. The CTD is required for linker stem formation [77] and the stabilization of secondary chromatin structures [75, 79]. The CTD is intrinsically disordered but becomes more structured and compact upon nucleosome binding [80]. However, the precise localization of the CTD on the nucleosome remains a major open question.

The linker histone globular domain adopts a winged-helix DNA-binding fold [81]. Over the past four decades, evidence has accumulated in support of distinct models for how the globular domain interacts with nucleosomes. According to one model, the globular domain is positioned on the nucleosome dyad and interacts with both DNA linkers [75, 77, 78, 82], whereas in alternative models the domain is displaced from the dyad and contacts only a single linker [83-85]. This debate has been significantly clarified by recent structural work, which demonstrated distinct binding modes for different linker histone-nucleosome complexes. Specifically, the crystal structure of the isolated globular domain of chicken H5 (GH5) bound to a 167-bp particle exhibits an on-dyad binding mode [86]. In contrast, an NMR study reported an off-dyad binding mode for the *Drosophila* H1 globular domain (GH1) [87]. The different binding modes utilized by these two histone isoforms have been ascribed to differences in a small number of DNA-contacting residues [88]. A distinct off-dyad binding mode was reported for human histone H1.4 in the cryo-EM structure of a condensed 12-nucleosome array [89]. Taken together, these studies suggest that the orientation of the linker histone globular domain may be isoform- and context-dependent. Consistent with this idea, a computational study identified three distinct, energetically favourable binding sites for the globular domain on the nucleosome [90].

RESULTS

MAPPING THE INITIATOR BINDING TAF2 SUBUNIT IN THE STRUCTURE OF HYDRATED YEAST TFIID

The studies described in this report were aimed at elucidating the location of Taf2p within the yeast TFIID (yTFIID) complex. Taf2p is the only integral TFIID subunit whose location within the complex has yet to be determined [91]. Given the large size (M_r=161 kDa) and key functionalities of Taf2p in promoter recognition noted above, there is a clear need to map the location of this subunit in the TFIID holocomplex. The complexity of TFIID has prevented structure determination by X-ray crystallography, and thus far electron microscopy has provided the only available structural models [92, 93]. Image analysis of isolated yTFIID molecules revealed a three-dimensional (3-D) model at 32Å resolution [3]. Human TFIID (hTFIID) was recently investigated in cryo electron microscopy but the resolution was not dramatically improved over that obtainied for yTFIID [2]. The authors described considerable conformational flexibility within hTFIID, which is a major limitation for high-resolution structural determination.

In the present work the structural variability of a TFIID preparation partially depleted in Taf2p was studied. This analysis allowed us to discriminate between two major contributions to structural variations: the conformational flexibility of the TFIID molecule and the lack of Taf2p. An additional domain was identified in some yTFIID molecules, and immuno-labeling demonstrated that this domain mapped to the N-terminal portion of Taf2p. The N-terminus of Taf2p bears sequence homology with the M1 family of metallopeptidases [94] whose atomic structure has been solved. This homology domain could be fitted into the additional density generated by Taf2p. The Taf2p content thus appears as a major source of specimen heterogeneity and the selection of a homogeneous Taf2p-containing TFIID subpopulation allowed us to significantly improve the resolution of the TFIID model to 23Å.

Taf2p Heterogeneity in TFIID Prepared by Different Procedures

During the course of pilot experiments to explore alternative purification strategies for yeast TFIID we observed large variations in the stoichiometry of the Taf2p subunit relative to our standard TFIID preparation. Our standard TFIID purification method used a yeast strain expressing an N-terminally HA_1-tagged Taf1p and involved Bio-Rex 70, anti-HA mAb and Mono-S FPLC chromatography [4, 95]. To investigate the use of the Tandem Affinity Purification (TAP) method for TFIID purification, we utilized a yeast strain expressing C-terminally TAP-tagged Taf1p [96]. The TAP method was optimized for TFIID solubilization and overall yield. We often observed reduced amounts of Taf2p relative to Taf1p in the TAP-purified TFIID (see Sypro Ruby stained SDS-PAGE, **Figure 1A).** Quantitation of the amounts of these two Tafs in the various preparations confirmed this observation; the Taf2p/Taf1p content of HA-TFIID was 0.7 compared to TAP-TFIID preparations A (Taf2/Taf1p = 0.1) and B (Taf2/Taf1p = 0.4).It is likely that Taf2p is preferentially lost from the TFIID complex. Such loss of the full-length Taf2p subunit was observed previously by Smale and colleagues who have reported that human Taf2p, which they termed CIF, readily dissociates from hTFIID [97].We have observed a similar phenomenon; a fraction of yeast Taf2p dissociated from yeast TFIID during ion exchange chromatography, perhaps in association with other TFIID subunits (**Figure 1B;** compare Taf2p/Taf1p amounts (asterisks) in the Mono-S Input (In) versus amounts in gradient Fractions, 4, 5-10). Together these results suggest that complete loss of the Taf2p subunit is the most likely explanation for the reduced stoichiometry of this subunit in the case of the TAP-purified TFIID.

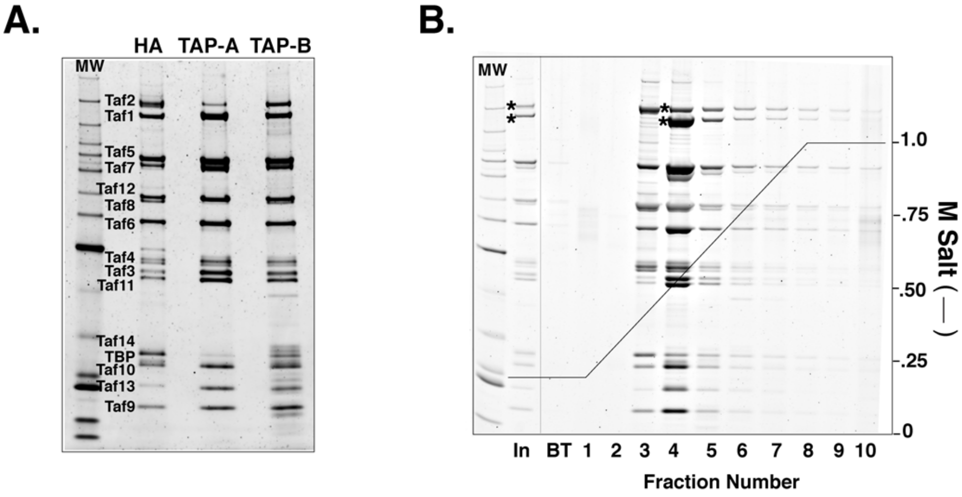


Figure 1 Heterogeneity in Taf2p content in TFIID purified by different methods.
(A) Sypro Ruby stained SDS-PAGE gel showing the subunit content of either HA-tagged Taf1p (HA) or TAP-tagged Taf1p (two independent preparations, TAP-A and TAP-B) purified TFIID. MW standards were run in parallel (MW) and the TFIID subunits are labeled. The stained gel was scanned (BioRad FX imager) and Taf1p and Taf2p content determined using QuantityOne software (BioRad). (B) Apparent dissociation of a portion of Taf2p from the TFIID holocomplex during Mono-S FPLC chromatography. CaM-Sepharose eluted TAP-TFIID was subjected to Mono-S FPLC chromatography and eluted with 1M NaCl as shown. The protein composition of the fractionated TFIID was measured by SDS-PAGE as in panel A; Fraction number, Molecular weight standards (MW), Input to Mono-S column (In) and unbound proteins (BT) indicated. The Taf2p and Taf1p subunits in the Mono-S 1M salt-eluted fractions are indicated by asterisks.

Single particle tomography

We decided to use the observed Taf2p subunit deficiency to our advantage given that otherwise the TFIID subunit composition of the TAP-TFIID was normal (*cf.* **Figure 1A**). We reasoned that image analysis of the TFIID preparation partially depleted in Taf2p should reveal different particle populations and identify the location of the subunit within the holocomplex. The TAP-A-TFIID preparation is likely to contain two major forms of structural variation: biochemical differences in Taf2p content and conformational changes. Standard image analysis of molecular views is unable to resolve such complex variation since most views cannot be unambiguously attributed to a given species. Consequently we used electron tomography of single particles to experimentally reveal these different TFIID populations. This method generates a low resolution 3-D model for each individual particle. A total of 157 individual molecular volumes were reconstructed from a field of negatively stained TAP-A-TFIID molecules. These volumes were aligned one with respect to the others using a combination of interactive and correlation based methods. The aligned volumes were classified into 28 groups according to maximal interclass resemblance and an average volume was calculated for each class. As a consequence of the limited resolution (estimated to be 50 Å on average) only large 3-D variations could be detected (**Figure**

2A). In general the volumes showed a three-lobed organization as reported previously, although 4 out of the 28 volumes apparently lacked one lobe, and the relative size of the lobes was found to be highly variable. The enclosed angle between the three lobes and the distance between the centers of the external lobes A and B also showed large variations. In the most closed conformation the distance between the lobes is 100 Å and the enclosed angle is 53˚, whereas in the most open conformation the distance between lobe A and B is 220 Å and the enclosed angle is 115˚.

The volumes obtained by electron tomography of single molecules provide a repertoire of the most extreme conformations that can be found in the TAP-A-TFIID preparation. The observed variability is likely to reflect the full range of conformational space as well as the biochemical variations of the data set plus any possible structural perturbations introduced by the preparation method.

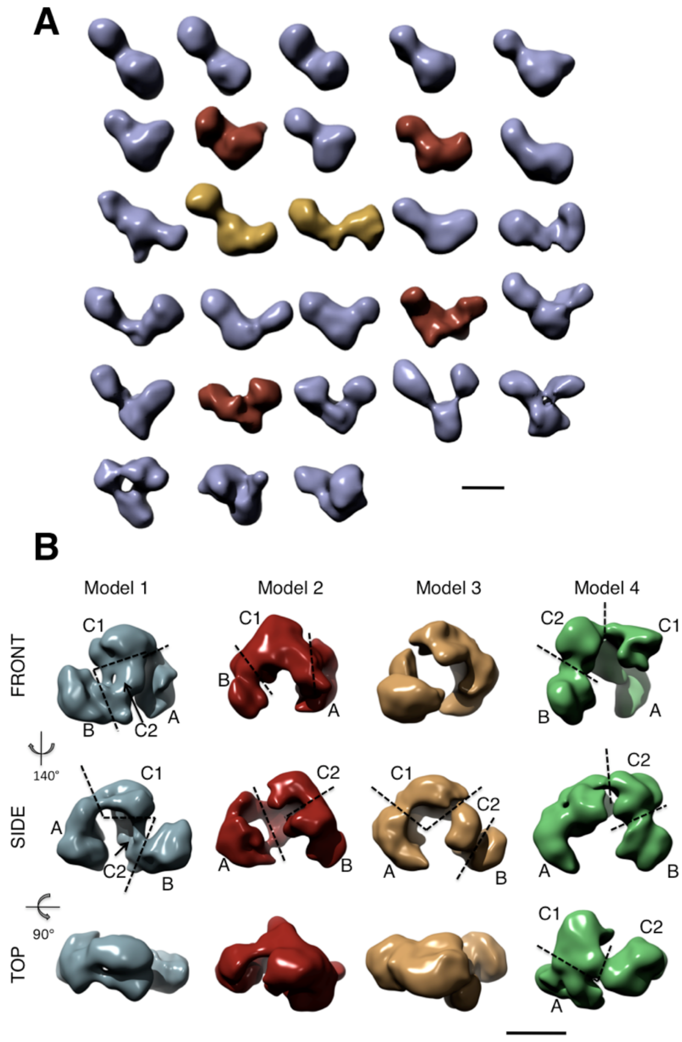


Figure 2 Structural variants of TFIID.
(A) Gallery of TFIID volumes obtained by electron tomography of negatively stained particles. The 28 represented volumes are averages obtained upon clustering of a total of 157 aligned individual volumes. The six volumes that were most frequently used in the refinement are colored differently; those which were progressively less used in the later refinement and discarded are in yellow whereas the four major variants are in red. (B) Four major structural variants obtained upon analysis of an image dataset of frozen hydrated, negatively stained TFIID molecules. All four models show three-lobes (A, B and C) and lobe C is separated into two domains (C1 and C2). The dashed lines represent the limits of each domain. The bars represent 10 nm.

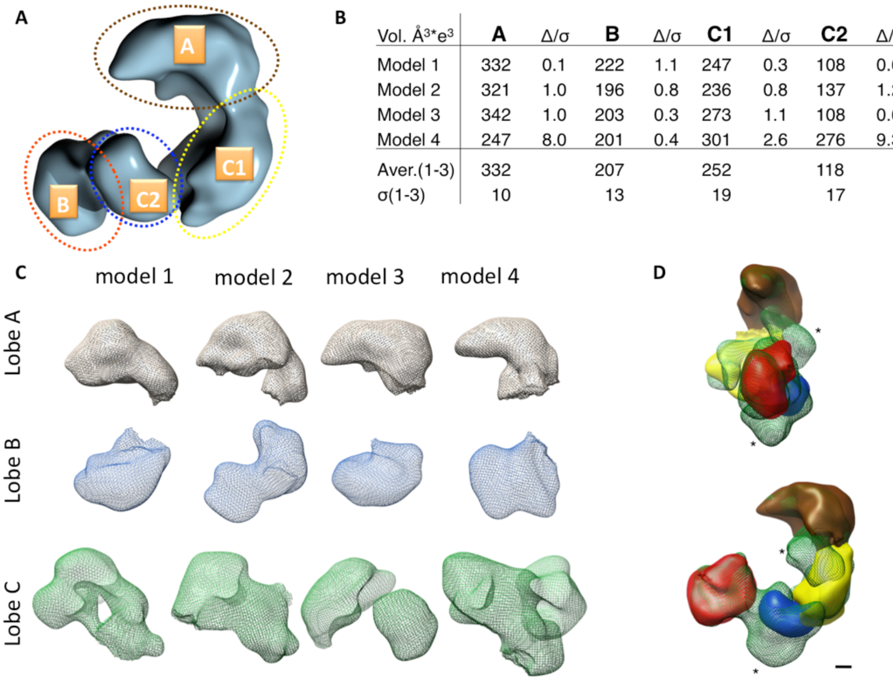
Analysis of the TFIID structural variants in negatively stained, frozen hydrated samples

In order to reveal greater structural detail, and to improve the statistical significance of the tomography volumes, we analyzed a large image dataset of 44,233 single particles for which two parameters were modified to reduce preparation-induced specimen variability. Firstly the TFIID molecules were observed under negative stain cryo-electron microscopy (NScryoEM) conditions in order to preserve the hydrated structure of the particle [98]. Secondly the molecules were chemically cross-linked according to the GraFix method [99] in order to lock the complex in a reduced number of conformational states. The tomographic volumes were utilized to sort the particle images into different categories and to verify which structural variants were present. Among the 28 average tomographic volumes, only 6 were used extensively since 75% of the images aligned with highest correlation against references issued from these volumes. The remaining 22 volumes were each used by less than 5% of the images and were discarded from further analysis. Upon refinement two of the six models were progressively less used and were discarded once they attracted less than 5% of the images. Four stable TFIID populations were thus identified, and a 3-D model was reconstituted for each (**Figure 2B**). These results suggest that dehydration is likely to be the source of a large part of the variation found in the tomographic volumes, and that only a limited fraction of the particles adopt the most extreme conformations after chemical cross-linking.

The four remaining models are organized as a molecular clamp formed by three or four successive lobes (A, B, C1 and C2) but differ significantly in the relative position of the lobes since the distance between the centers of the external lobes varied from 140 Å to 159 Å. When the same intensity threshold is set to all models, model 4 was found to be larger than the three other models. Models 1, 2 and 3 comprise an average volume of 910 nm^3 (911, 891 and 927 nm^3, respectively) whereas the volume of model 4 is 1027 nm^3, a value 6.5σ above the average volume of the other models. This indicates that the TFIID molecules contributing to model 4 enclose an additional mass. The direct superposition of the models however, was not possible because of the complex structural transition experienced by TFIID (**Figure 2B**). The 3-D models were therefore split into four sub-domains that were consistently observed in order to align the domains one with respect to the others and to analyze the TFIID rearrangements more accurately (**Figure 3A**). The A-lobes issued from models 1-3 have a similar beak-like shape of 10.5 by 6 by 6 nm in size and occupy an average volume 332 nm^3 (σ = 3.2%) whereas in model 4 the A-lobe appears slimmer and occupies a volume of 247 nm^3 (**Figure 3C**). The volume of the B-lobes was stable in all models (206 nm^3, σ = 5.6%) and their shape was also fairly similar except for model 2 in which it appeared split into two sub-domains and thus more elongated. Lobe C showed the largest variability in size and shape and could be divided into two modules, C1 and C2, whose relative orientations vary, suggesting that these modules can move as independent entities. Whereas the C1- and C2-lobes from models 1-3 were similar in size with an average volume of 252 nm^3 (σ = 7.6%) and 118 nm^3 (σ = 14.2%) respectively, these two lobes were larger in the case of model 4. While the size of lobe C1 showed a moderate increase of 49 nm^3 (2.6 σ above the average of models 1-3), the volume of lobe C2 more than doubled with an increase of 158 nm^3 (9.3 σ above the average of models 1-3).

Altogether, these results indicate that model 4 comprises an additional protein density in lobe C2. In order to better delineate the additional density and to compensate for the flexibility of the structure, the dissected lobes from model 3 were individually positioned into model 4. This flexible domain docking revealed two additional protuberances present only in model 4: a large one in lobe C2 and a smaller one that connects lobe C1 to the A-lobe (indicated by asterisks in **Figure 3D**).

In order to assess whether the structural heterogeneity found in the TAP-A-TFIID preparation is related to the Taf2p depletion detected by the biochemical data, we analyzed the relative abundance of the four major TFIID populations within an image dataset of the HA-TFIID preparation that contains a higher Taf2p complement. Model 4, which represents 23% of the TAP-A-TFIID preparation, increased to 55.2% in the HA-TFIID preparation indicating that the largest TFIID structure is more abundant in the preparation that contains more Taf2p thus confirming that the additional protein domain found in model 4 reflects the presence of this subunit. Model 4 is thus the most comprehensive TFIID structure and was therefore further refined to reach the model shown in **Figure 6**. The resolution tests for this model calculated from 10,205 images gave values of 23 Å and 19 Å for the 0.5 Fourier shell correlation and the half-bit criteria, respectively. Interestingly the size of the additional domain located between the C1 and A lobes increased significantly upon refinement.



Vol. Å³*e³	A	Δ/σ	B	Δ/σ	C1	Δ/σ	C2	Δ/σ
Model 1	332	0.1	222	1.1	247	0.3	108	0.6
Model 2	321	1.0	196	0.8	236	0.8	137	1.2
Model 3	342	1.0	203	0.3	273	1.1	108	0.6
Model 4	247	8.0	201	0.4	301	2.6	276	9.3
Aver.(1-3)	332		207		252		118	
σ(1-3)	10		13		19		17	

Figure 3 Comparison of the obtained TFIID models.

(A) Dissection of TFIID model 3 into 4 sub-domains A, B, C1 and C2. (B) Table showing the volume occupied by the different lobes A, B, C1 and C2. The average volume (Average) of a given lobe as well as the standard deviation (?) are shown in the last rows. Note that for lobe C2 these values were calculated from models 1, 2 and 3 only. The columns labeled "Δ/ ?" represent the difference between the volume of a lobe and the average volume of that lobe normalized by the associated standard deviation. (C) Alignment of lobes A, B and C dissected from all four TFIID models showing their structural homology. (D) Fitting of TFIID model 3 into the envelope of the larger model 4 reveals two additional densities (highlighted by asterisks; *) in lobe C; colors, red, blue, yellow and brown represent lobes B, C2, C1 and A, respectively. Bar indicates 2 nm, model bottom was rotated 97° relative to top model

Antibody labeling of Taf2p

To further confirm that the additional density in model 4 corresponded to Taf2p, HA-TFIID was labeled with an antibody raised against a peptide corresponding to residues 5-19 of Taf2p. The specific immune complexes formed upon incubating the peptide affinity-purified antibodies with TFIID were negatively stained and visualized by electron microscopy. A total of 1,600 individual TFIID-IgG images were selected and aligned against references issued from model 4, which we hypothesized contained the full length Taf2p. After image clustering, the class-averages where the antibody was clearly bound to TFIID were used to calculate a 3-D map in which the position of the antibody binding site could be determined by density difference with the unlabeled model 4 (**Figure 4A**). The bound antibody highlighted the additional density present in the C2-lobe of model 4, a result consistent with the hypothesis that this density defines the Taf2p subunit of TFIID.

The peptide used to generate the antibody overlaps with the first six residues of the leukotriene A4 hydrolase homology domain. This homology extends over Taf2p residues 11–533, and the entire conserved catalytic region defining the M1 protease family (residues 1–458 of leukotriene A4 hydrolase). This region comprises structural domains A and B in the crystal structure of the protease [100]. The homology between Taf2p and leukotriene A4 hydrolase is substantial since overall, 18% of the residues are identical for the two domains. Moreover, the pattern of hydrophobic core residues of the protease appears well conserved while the identity of important residues in turns (predominantly Pro and Gly) and numerous acidic and basic residues that form intramolecular salt bridges in the crystal structure have also been conserved between the two proteins (**Figure 4B**). Taken together, these observations indicate that Taf2p is likely to adopt a ternary structure highly similar to the A and B domains of leukotriene A4 hydrolase. No significant homology exists between the remainder of the Taf2p sequence and the additional domain in the crystal structure, which is specific to the leukotriene A4 hydrolase subfamily of the M1 proteases. The docking of the atomic structure indicated that the additional density revealed in model 4 has a size large enough to accommodate the human Leukotriene A4 hydrolase domain (**Figure 4C**). Thus the slightly elongated and asymmetric shape of the domain fits within the external contours of the C2 domain envelope, thereby providing additional constraints for a precise positioning of the atomic structure. Together these experiments indicate that the N-terminal part of Taf2p is located in lobe C2 of TFIID.

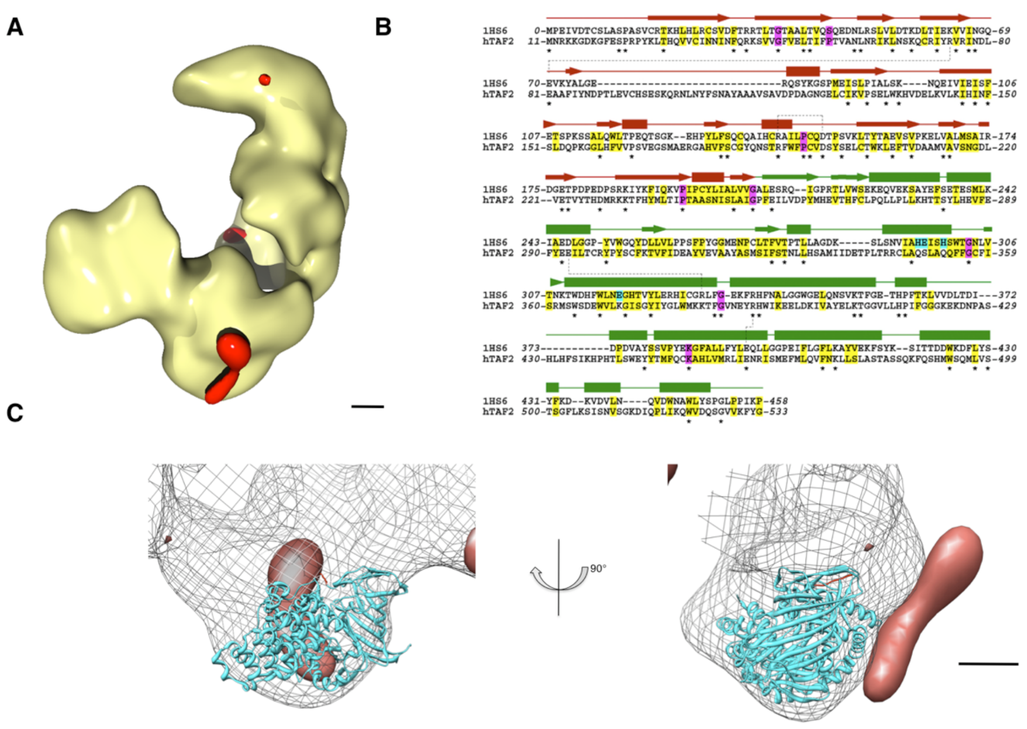

A
B
C
1HS6 0-MPEIVDTCSLASPASVCRTKHLHLRCSVDFTRRTLTGTAALTVQSQEDNLRSLVLDTKDLTIEKVVINGQ-69
hTAF2 11-MNRKKGDKGFESPRPYKLTHQVVCINNINFQRKSVVGFVELTIFPTVANLNRIKLNSKQCRIYRVRINDL-80
1HS6 70-EVKYALGE----------------------------RQSYKGSPMEISLPIALSK----NQEIVIEISF-106
hTAF2 81-EAAFIYNDPTLEVCHSESKQRNLNYFSNAYAAAVSAVDPDAGNGELCIKVPSELWKHVDELKVLKIHINF-150
1HS6 107-ETSPKSSALQWLTPEQTSGK--EHPYLFSQCQAIHCRAILPCQDTPSVKLTYTAEVSVPKELVALMSAIR-174
hTAF2 151-SLDQPKGGLHFVVPSVEGSMAERGAHVFSCGYQNSTRFWFPCVDSYSELCTWKLEFTVDAAMVAVSNGDL-220
1HS6 175-DGETPDPEDPSRKIYKFIQKVPIPCYLIALVVGALESRQ--IGPRTLVWSEKEQVEKSAYEFSETESMLK-242
hTAF2 221--VETVYTHDMRKKTFHYMLTIPTAASNISLAIGPFEILVDPYMHEVTHFCLPQLLPLLKHTTSYLHEVFE-289
1HS6 243-IAEDLGGP-YVWGQYDLLVLPPSFPYGGMENPCLTFVTPTLLAGDK-----SLSNVIAHEISHSWTGNLV-306
hTAF2 290-FYEEILTCRYPYSCFKTVFIDEAYVEVAAYASMSIFSTNLLHSAMIIDETPLTRRCLAQSLAQQFFGCFI-359
1HS6 307-TNKTWDHFWLNEGHTVYLERHICGRLFG-EKFRHFNALGGWGELQNSVKTFGE-THPFTKLVVDLTDI---372
hTAF2 360-SRMSWSDEWVLKGISGYIYGLWMKKTFGVNEYRHWIKEELDKIVAYELKTGGVLLHPIFGGKEKDNPAS-429
1HS6 373-----------DPDVAYSSVPYEKGFALLFYLEQLLGGPEIFLGFLKAYVEKFSYK--SITTDDWKDFLYS-430
hTAF2 430-HLHFSIKHPHTLSWEYYTMFQCKAHLVMRLIENRISMEFMLQVFNKLLSLASTASSQKFQSHMWSQMLVS-499
1HS6 431-YFKD--KVDVLN----QVDWNAWLYSPGLPPIKP-458
hTAF2 500-TSGFLKSISNVSGKDIQPLIKQWVDQSGVVKFYG-533
90°

Figure 4 TFIID immunolabeling with Taf2 antibody
(A) Immunolabeling of TFIID by an antibody directed against the N-terminus of Taf2p. The 3-D model of TFIID is represented in yellow and the red surface represents the difference map between the antibody-labeled TFIID and the unlabeled complex. (B) Alignment between Leukotriene A4 hydrolase and human Taf2p. The conserved amino acids are highlighted. (C) Docking of the atomic structure of the leukotriene A4 hydrolase domain that is homologous to the N-terminal part of Taf2p into the additional density present in lobe C2 of model 4 shown in teal. The bars represent 2 nm; indicated in red is the additional density of the anti-Taf2p IgG.

Plasticity of the TFIID structure

The comparison between the four TFIID models indicates that with the exception of lobe C2, the domains are mostly conserved in size and shape, but that the overall structure of TFIID is variable. To further describe these TFIID domain movements, each NScryoEM model was represented as a skeleton where the center of each lobe is schematized by a disk, each connected by a wire (**Figure 5A**). Angle α was defined as the angle between lobes A, C1 and C2, and β as the angle between lobes C1, C2 and B. The variation in α appears continuous within a range of 60°, from 53˚ (model 1) to 113˚ (model 4) (**Figure 5B**). The angle β shows a smaller variation (23°) and appears to adopt two values: 120° in models 1 and 4 and 97° in models 2 and 3 (**Figure 5C**). This suggests that lobe B can adopt two discrete conformations relative to lobe C. Another level of variability is the rotation of lobes around their connections. In this respect, lobes C2 and B show the highest flexibility since their relative orientation was found to change by up to 73˚ between their positions in models 1 and 2. This contrasts with the fixed position of lobe A relative to lobe C1. Collectively these data argue that the yTFIID complex is a flexible assembly that is likely capable of adopting a number of distinct conformations.

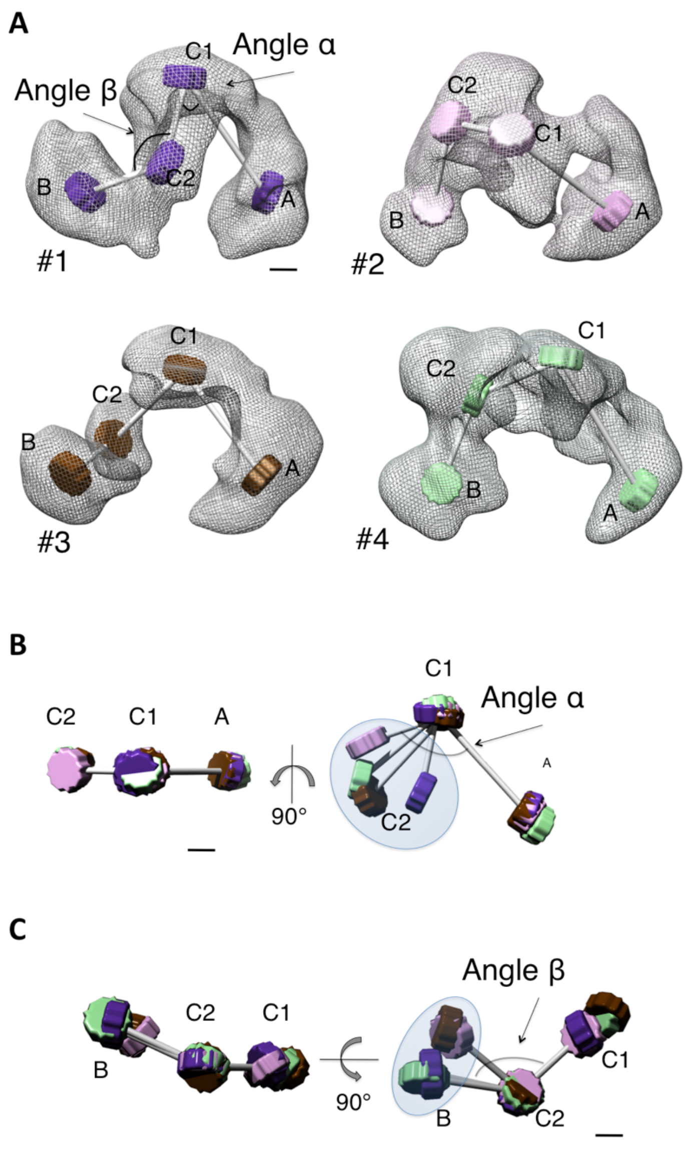


Figure 5 Flexibility of the TFIID structure.

(A) TFIID models 1, 2, 3 and 4 are represented as skeletons in which the centers of the four major domains are symbolized by disks and the domain connections by rods. (B) Variations in the ✔ angle between domains A, C1 and C2. (C) Variations in the 🚲 angle between domains B, C2 and C1. The bars represent 2 nm.

Discussion

The present cryo-EM study of frozen hydrated and negatively stained yeast TFIID molecules provides the highest resolution map available for this multiprotein transcription initiation factor and stresses the importance of specimen heterogeneity in the quest of finer details. Earlier attempts have previously shown that conformational changes constitute a major limitation for reaching high resolution structural information [2]. Our results further emphasize the importance of variations in subunit composition, which may, as in the case of Taf2p, affect specimen homogeneity. Interestingly our new model not only shows greater structural detail but also converges towards the structure of the human TFIID determined independently [2].

The structural characterization of molecular variations by electron microscopy can be resolved by either experimental or computational methods [101]. In all computational approaches a starting model is required either for alignment of the particles, or for angular assignment of the particle views for 3-D reconstruction. The use of a starting model may however bias the analysis, particularly if large structural fluctuations are suspected since alignment and/or angular assignment may be incorrect for the most extreme conformations. In this report specimen heterogeneity was addressed experimentally by performing electron tomography to generate a 3-D model for each molecule without the need for any alignment or clustering before 3-D reconstruction. The drawbacks of using this method, such as accumulated electron irradiation and large missing wedge of information that result from the tilting experiment were overcome, to some extent, by respectively lowering the electron dose and *a posteriori* averaging of molecular volumes. Further progress in resolution will need to address any additional structural and conformational fluctuations by biochemical improvement of specimen homogeneity, or stabilization of a particular conformation, by using more accurate tomographic data collection from cryo samples to detect subtle intermolecular variations and by developing computational methods to separate heterogeneous image datasets.

Conformational changes within yTFIID

The gallery of volumes resulting from the tomography experiments showed extremely large movements of the major domains. The conformational space explored by TFIID is considerably reduced when the hydrated state of the molecules is preserved and upon stabilization of a limited number of structural states by the GraFix method. Nevertheless, the refinement process selected out four abundant states of TFIID and revealed complex conformational transitions that are not limited to a spring-like flexibility between the most extreme lobes but involve significant reorganizations within domains. The four identified states could either represent stable conformations or average snapshots of a continuous structural transition. Two lines of evidence indicate that at least some transitions are not completely continuous and that stable states may exist. First, lobe B appears to adopt two discrete positions relative to the rest of the structure (angle β in **Figure 5**) while lobe C2 seems to adopt a continuous variation. Second, only one of the four states showed an additional density suggesting that the presence of Taf2p stabilizes a particular TFIID conformation.

Whereas lobes A and B appeared most stable, lobe C was found to undergo considerable conformational transitions. This observation is consistent with previous work that analyzed conformational variations in human TFIID, where it was also noted that the central domain is subject to reorganization [2]. The precise mechanism and the functional significance of these conformational changes remain to be elucidated. The possibility that the structure of TFIID can be adapted to allow for the recognition of a large variety of promoters, each with distinct activator binding site distributions is particularly attractive. Along these lines, the recently reported structural change of human TFIID upon incorporation of the cell-type specific Taf4b paralogue instead of Taf4p was correlated with modified promoter selectivity [102]. These observations suggest that the precise conformation of the TFIID complex may contribute to the specificity of promoter recognition. Alternatively, transcription factor-TFIID interaction(s) may alter TFIID conformation and directly modify promoter selectivity.

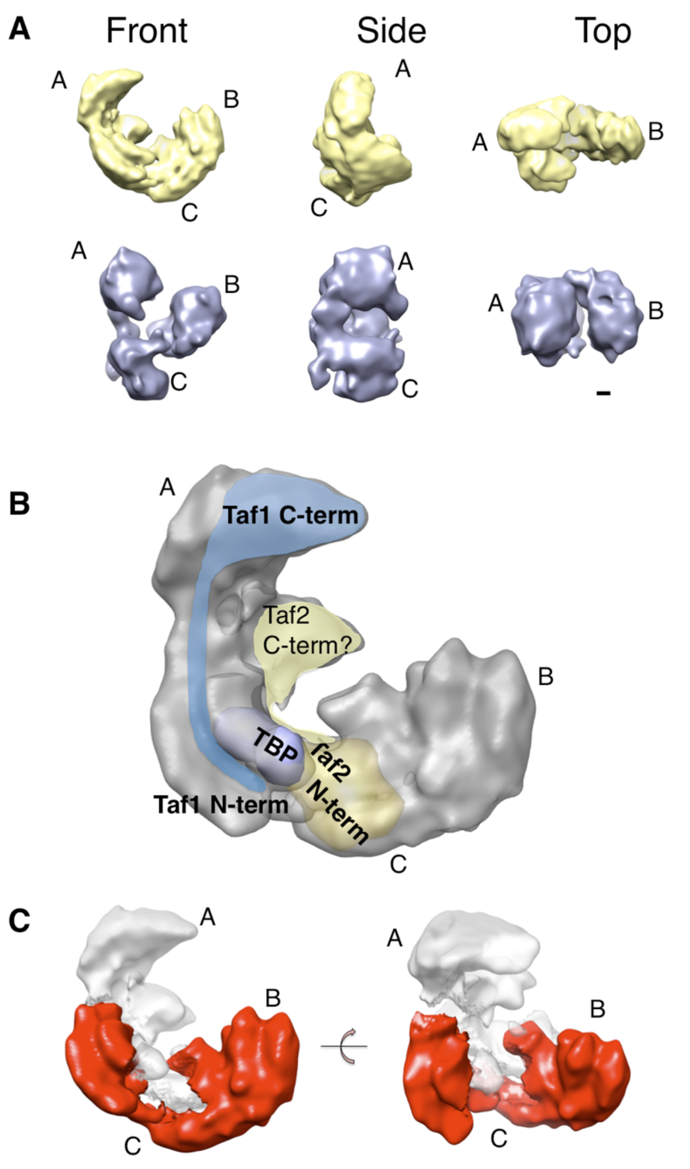


Figure 6 Molecular environment of Taf2p within TFIID

(A) Alignment of the previous TFIID volume (lower row; violet) with the higher resolution model described in this report (upper row; yellow); the three lobes of the structure A, B and C are labeled. (B) Localization of Taf2p relative to previously mapped Taf1p and TBP. (C) Identification of a crescent-shaped TFIID subcomplex obtained upon removal of the Taf1p, Taf2p and TBP densities.

Taf2p structure-function relationships

The localization of Taf2p was determined by taking advantage of biochemical variation in Taf2p content and by immuno-labeling. Collectively these approaches showed that the N-terminal part of Taf2p is located in the C-lobe. Moreover, this domain has a size and a shape capable of accommodating the homologous aminopeptidase fold. A second protein domain was also found to be missing in those TFIID particles that do not contain the N-terminal part of Taf2p. This domain connects lobes C to lobe A and may correspond to the C-terminal part of Taf2p. Taken together these data suggest that the complete subunit is missing from TFIID, and that the loss of Taf2p apparent by SDS-PAGE analysis of TAP-tagged TFIID is most likely not due to partial proteolytic cleavage.

In order to position Taf2p relative to the previously mapped TFIID subunits [3, 91], our former 32 Å resolution model was aligned against the present 23 Å resolution model and an unambiguous superposition could be obtained (**Figure 6A**). Our results indicate that the C-terminal part of Taf2p maps close to TBP, a TFIID subunit that interacts with the N-terminus of Taf1p with high affinity [52, 103, 104] **(Figure 6B)**. This proximity is consistent with earlier data showing a direct interaction between these three proteins forming a stable subcomplex capable of binding promoter DNA *in vitro* [105]. Although the existence of such a subcomplex has yet to be demonstrated *in vivo*, the finding that in the *P. falciparum* genome homologues of Taf1p, Taf2p and TBP, but of no other Tafs (with the possible exception of Taf10p) could be identified, strengthens the functional significance of a Taf1p-Taf2p-TBP subcomplex [94]. Within TFIID this ternary complex is likely to be extended and to encompass both lobes A and C since the largest C-terminal part of Taf1p was found to be located in lobe A where it could contact the C-terminal part of Taf2p, whereas the N-terminal end of Taf1p reached toward lobe C where the N-terminal part of Taf2p was mapped [91]. TBP was also found to be located between the A and C lobes. In this respect it is notable that a C-terminal 369 aa fragment of Drosophila Taf2p was reported to bind directly and independently to both TBP and Taf1p [103]. The major contacts between these polypeptides are thus likely to occur through the C-terminal portion of Taf2p, which we speculate corresponds to the protein density located between lobes A and C.

The extended DNA binding profile of TFIID compared to TBP can be partially mimicked by the Taf1p-Taf2p-TBP ternary complex suggesting that this subcomplex contains many of the DNA binding properties of TFIID [46, 103]. Taf2p by itself was reported to interact with the initiator core promoter element (Inr) [46, 97, 103] and in the context of human TFIID, Taf2p could be cross-linked to the Adenovirus Major Late Promoter Inr element [106]. Taken together these results indicate that Taf2p is likely to participate in start-site selection on certain promoters through its ability to bind the Inr. The relative positions of TBP and Taf2p in our TFIID model thus suggest an orientation of the promoter DNA: the upstream region should contact TBP whereas the transcription initiation site of the promoter would be positioned closer to the Taf2p region. The TBP-Taf2p location further defines a curved DNA binding interface that, according to our model, can be as large as 15 nm and could thus accommodate up to 45bp of promoter DNA within the clamp. It is tempting to speculate that this distance is related to the downstream DNase I hypersensitive sites which were detected on several yeast promoters at 45 bp downstream of the TATA box [4] but additional experiments are required to address this question. Finally the proposed location for the promoter DNA within the clamp formed by TFIID gives strong spatial constraints for the assembly of the pre-initiation complex. The volume forming the clamp is large enough to contain a sphere with a diameter of 12.5 nm and could thus accommodate several transcription proteins.

Identification of a TFIID core sub-complex

Several lines of evidence indicate that the 3-D architecture of TFIID is likely to contain a stable core subcomplex onto which an independent module is assembled. The use of RNAi to probe the stability of the TFIID complex in *Drosophila* tissue culture cells revealed the existence of a stable core-TFIID subcomplex composed of Taf5p and the two Histone Fold Domain (HFD) -containing Taf pairs Taf4/12 and Taf6/9 [107]. This core-TFIID is believed to be decorated with peripheral subunits, in particular those which compose the Taf1p-Taf2p-TBP subcomplex. Additional results consistent with these observations had previously been reported during biochemical analyses of TFIID purified from yeast expressing a temperature-conditional mutated form of Taf1p [108]. In order to translate these observations onto our electron microscopy map, the potential protein densities of Taf1p, Taf2p and TBP were removed from the TFIID envelope. The shape of the remaining structure is reminiscent of that of a stable *in vitro* reconstituted complex composed of Taf5p and the three HFD-pairs Taf4/12, Taf6/9 Taf8/10 [91] and interestingly presents an almost symmetric crescent-shaped structure (**Figure 6C**). Consistent with this observation are the biochemical quantization and *in vivo* self-association properties of yeast Tafs which showed that several polypeptides are present with more than one copy in each TFIID molecule [4]. These results were confirmed by immuno-electron microscopy data that showed that at least Taf5p and the five HFD-containing subunits (Taf10p, Taf6p/9p, Taf4p/12p) are present as two copies [3]. The Tafs present as two copies are likely to form a two-fold symmetric assembly similar to the subcomplex outlined within TFIID. Altogether, the higher resolution structure and the immuno-labeling studies strongly suggest that TFIID is composed of two subcomplexes. On one hand a core complex containing Taf5p and most of the HFD-containing Taf-pairs (Taf6/9, Taf4/12, Taf8/10, Taf11/13) is proposed to adopt a crescent-shaped two-fold symmetric structure. On the other hand a complex containing Taf1p, Taf2p and TBP, as well as probably Taf7p, is predicted to be recruited to this core complex.

In summary, we have mapped the location of the Taf2p subunit within the yeast TFIID holocomplex. Further, by capitalizing on the sub-stoichiometric Taf2p content of certain preparations of TFIID, complemented by single particle tomography, immuno-labeling and cryo EM, we have generated a 23 Å model of yeast TFIID. This new model has both higher structural resolution as well as increased definition of the location of the three TFIID subunits, Taf1p, Taf2p and TBP that likely participate critically in both promoter binding and promoter selectivity. Additional experimentation designed to provide further details of yeast TFIID structure, both alone and complexed with other GTFs, promoter DNA and transactivator proteins are in progress. This work will provide additional insights into how TFIID subserves both its coactivator and promoter recognition functions.

Materials and Methods

TFIID Purification. HA_1-Taf1p-tagged TFIID was purified from *Saccharomyces cerevisiae* strain YSLS18 as described previously [4, 95]; note that this purification scheme utilized ethidium bromide to prevent TFIID-DNA interactions during purification. TAP-Taf1p-tagged TFIID was purified from yeast strain YLSTAF1 (kindly provided by Dr. Ray Jacobson, University of Texas MD Anderson Cancer Center). This strain expresses Taf1p with 4.5 copies of the TAP tag (Protein $A_{4.5X}$-TEV protease site-Calmodulin Binding Domain; [109]) at the C-terminus. YLSTAF1 cells were grown to mid-log phase, harvested and processed for TFIID purification as for YSLS18. An overview of the three steps used for the subsequent purification of TFIID are as follows: (a) solubilized TAP-Taf1p TFIID was bound to IgG Sepharose and eluted using the TEV protease, (b) the IgG Sepharose eluate was bound to Calmodulin (CaM)-Sepharose and eluted with EGTA, (c) the CaM-Sepharose eluates were immediately loaded onto a MonoS FPLC column, the column washed, and TFIID eluted with a gradient of 1M salt. The TAP-tagged TFIID prepared by this procedure is highly concentrated (1-3 mg/ml); typical yield 3-3.5 mg TFIID/kg cells. Neither preparation generated TFIID with significant amounts of either contaminating DNA or RNA (**Figure S5**).

Anti-Taf2p Peptide Antibodies. Rabbit antibodies against Taf2p N-terminal amino acid residues 5-19 (SKNATPRAIVSESST) were prepared by Antagene Inc (Mountain View, CA). Peptide immobilized on Sulfolink beads (Pierce, Inc.) with an added C-terminal cysteine residue was used for affinity purification of antibodies specifically recognizing Taf2p N-terminal sequences. These antibodies only recognized the Taf2p subunit of TFIID (**Figure S6**).

Electron tomography. Single tilt tomography was performed at room temperature using a FEI Tecnai F20 electron microscope operating at 200 kV. Specimens were observed under low-dose conditions (total dose of 40-50 e/Å$^{2)}$ with a tilt range of -65 to +65 degrees. Specimens were sandwiched between two layers of carbon using 2% uranyl-acetate as a stain and 0.1% glutaraldehyde as a cross-linking agent. The tomograms were reconstructed using the IMOD software package [110]. To investigate the differences between the 3-D models the maps were first roughly aligned in real space using Chimera's "fit map in map" tool [111]. This pre-alignment was refined by cross correlation and the aligned volumes were clustered after multivariate statistical analysis using the IMAGIC software package [112].

Single particle cryo electron microscopy. The yTFIID sample was prepared using the GraFix method [99] in a buffer containing 10 mM Tris-HCl (pH 8.0), 300 mM NaCl and according to the cryo-negative stain method [98]. Images were collected at liquid nitrogen temperature under low-dose condition (15-20 e/Å^{2}), at a magnification of x40,000 on Kodak SO-163 films. The defocus values ranged from 0.137 to 1.96 µm. Micrographs with no visible drift were digitized with a 5µm raster size using a drum scanner (Primescan D7100, Heidelberg) and coarsened twice to obtain a final pixel spacing of 0.254 nm. Boxing and CTF phase flipping of the 44,233 TFIID images were performed in the EMAN software package. The image processing was performed using the IMAGIC (Image Science Software, Berlin, Germany) and Spider [113] software packages as described earlier. The resolutions of the final reconstructions were estimated according to the 0.5 cut-off in the Fourier shell correlation curve (0.5 FSC criterion) and the intersection point of the half bit curve with the FSC curve (half bit criterion) [114]. The final reconstructions were filtered to the measured resolution.

Immuno-electron microscopy. For immuno-electron microscopy a 3- to 5-fold molar excess of anti-Taf2p IgG was incubated 30 min at 20°C with purified HA-tagged TFIID at a final protein concentration of 30 µg/ml. Images of TFIID molecules putatively labeled by the IgG were collected and aligned against projections of model 4 (see **Figure 3B**). The aligned images were then analyzed by using multivariate statistical methods and hierarchic ascendant classification. Class average images were selected where the antibody bound to TFIID was clearly recognized and a 3-D map of the complex was determined to position the antibody binding site. The obtained volume along model 4 was normalized and a difference map was created by subtracting model 4.

TFIIA AND THE TRANSACTIVATOR RAP1 COOPERATE TO COMMIT TFIID FOR TRANSCRIPTION INITIATION

Results

Three-dimensional (3-D) electron microscopy (EM) has provided structural models for yeast TFIID [1, 115] that bears close similarities with its human ortholog [2, 3]. TFIID can be divided into five modules [1, 2], lobes A, B, C1, C2 and D that adopt a clamp-like structure in which lobes A and B form the jaws (Fig. 1a). TFIID serves as a coactivator for yeast transcription factor Rap1 through direct interactions with TFIID subunits Taf4, 5 and 12. The Rap1 transactivator was used here as a model to study the architecture of a DNA-bound activator-TFIID-TFIIA complex [42]. Rap1 is a multifunctional protein that plays important roles in gene transcription and telomere length regulation, acts both as a repressor or activator on different genes [116], drives transcription of over 40% of Pol II transcription in yeast, and is essential for transcription of the ribosomal protein genes used here as a model [42, 117]. In order to elucidate the architecture of an activator-TFIID-promoter DNA complex we assembled three TFIID-containing complexes and solved their molecular structure to a resolution of 18.6 - 35 Å, mostly by utilizing cryo-EM methods.

To visualize the Rap1 binding site on TFIID we first analyzed the structure of the TFIID-Rap1 complex formed as described in Methods with five-fold molar excess of Rap1. The recorded image dataset was analyzed and ultimately separated into two species, Rap1-TFIID complex, and holo-TFIID. A comparison of the two resulting models revealed that while the TFIID-Rap1 complex contained an additional density, the binding of Rap1 did not induce major conformational changes in TFIID, in agreement with a recent report describing the interaction of human TFIID with activators [118]. The additional Rap1-dependent protein density is found in TFIID lobe B (Fig. 1b) where it colocalizes with Taf5 and the Histone Fold Domain containing heterodimer Taf4/Taf12, consistent with our immunolabelling experiments [3, 91] and with Rap1-Taf binding studies [42, 119], thus we conclude that the extra mass in the TFIID-Rap1 complex corresponds to Rap1 binding with Tafs 4, 5 and 12.

In order to position TFIIA and TFIID-DNA interactions in the absence of activator we recorded cryo-EM images of a TFIID-TFIIA-DNA complex reconstituted on the Adenovirus 2 Major Late Promoter (Ad2 MLP) [4] with forty-fold molar excess of TFIIA and four-fold molar excess of DNA. Upon refinement, the image dataset was sorted to remove free, non-DNA bound TFIID molecules thus yielding 3-D models for the TFIID-TFIIA-DNA complex and holo-TFIID (**Fig. 7c**). In the TFIID-TFIIA-DNA complex an additional globular density was bound to lobe C1, which showed a rod-shaped extension towards lobe D. The globular density is compatible with the size of TFIIA and positions close to the mapped location of TBP [91] consistent with the known TBP-TFIIA interaction [120, 121]**.** The rod-shaped density connects to lobe D and runs along its surface where we have mapped the C-terminus of Taf2. Since Taf2 is documented to interact with the Inr sequence at the transcription start site in the Ad2 MLP [105, 106], we propose that the extra density depicted in green in **Fig. 7c** corresponds to promoter DNA. The additional mass present in the TFIIA-TFIID-DNA complex thus contains TFIIA and a promoter DNA fragment connecting TBP to Taf2.

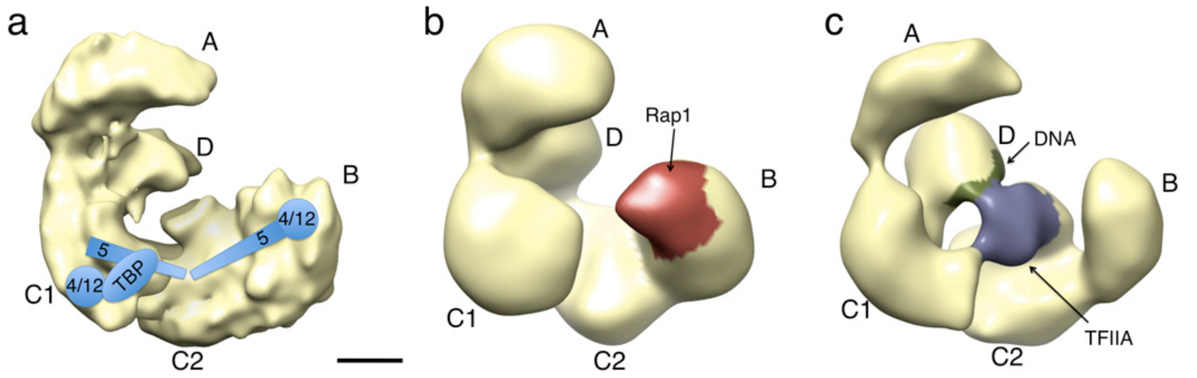


Figure 7 Location of critical components of the initiation process within various TFIID complexes. **a**, Cryo-EM structure of the yeast holo-TFIID complex. The five major lobes (A, B, C1, C2 and D) [1, 2] are depicted along with the location of TBP, Taf4, 5 and 12. Taf5 and the histone-fold Tafs, including Taf4 and Taf12, are present in two copies in yeast TFIID [3, 4] forming a crescent-shaped complex with two-fold symmetry [1]. **b**, Negatively stained structure of the TFIID-Rap1 complex. The additional density corresponding to Rap1 is colored in red according to difference maps shown in Supplementary Fig. 5a. **c**, Cryo-EM model of the unstained TFIID-TFIIA-DNA complex formed between TFIID, TFIIA and the Ad2 MLP. Additional densities present in the TFIID-TFIIA-DNA complex are colored. The mass corresponding to TFIIA is represented in blue whereas the density arising in the D lobe ascribed to DNA is represented in green.

We next incubated TFIID, TFIIA and Rap1 with a chimeric yeast enhancer-promoter DNA fragment composed of the 41 bp Rap1 Upstream Activating Sequence of the ribosomal protein gene *RPS8A* (*UAS$_{RAP1}$*) fused to the *PGK1 core* promoter *[122]; **UAS$_{RAP1}$-PGK1***, by using a ten-fold molar excess of the DNA fragment and five-fold excess of both TFIIA and Rap1. Importantly, this minimal construct which contains two Rap1 binding sites fused to the *PGK1* promoter, has been well characterized. Transcription of this construct is both TFIID- and Rap1-dependent *in vivo* and *in vitro* [42, 122]. A large dataset of 110,000 images was collected, aligned and classified according to particle orientation. Substantial heterogeneity was observed for similarly oriented particles indicating that the dataset contained a mixed population of complexes. Therefore 3-D multivariate statistical analysis was applied to separate different TFIID complexes [123]. This method reveals three dominant structures termed Complexes I, II and III. Complex III corresponds to free, unbound holo-TFIID, and serves here as an internal reference structure.

Complex I contains TFIID, Rap1 and enhancer promoter DNA and is characterized by two additional densities bound to both faces of Lobe B (red, Fig 2a) positioned similarly to Rap1 within the TFIID-Rap1 complex. Importantly, the shape of the inner density, closest to lobe C1, accommodates the crystal structure of the Rap1 DNA Binding Domain (DBD) [124] (**Fig**. **8b**). This observation, indicating that Rap1 DBD binds directly to Lobe B, is consistent with biochemical and genetic evidence demonstrating that this central domain of Rap1 (residues 361 to 596) interacts with TFIID [42]. Moreover, this density shows a rod-shaped protuberance running towards and contacting lobe C2 (Lobe C2, green in **Fig**. **8a**). When the atomic structure of the Rap1-DBD-DNA complex [124] is fitted into the inner density the orientation and position of DNA overlaps the protruding rod suggesting that the narrow density connecting Rap1 on lobe B to lobe C2 corresponds to promoter DNA downstream of the Rap1-binding site. The shape of the density bound to the external face of lobe B, opposite to lobe D, is different from that of the Rap1-DBD and there is no protruding rod that could account for bound DNA, suggesting that it corresponds to other domains of Rap1. The 92 kDa Rap1 consists of DNA bending, BRCT, DBD, toxicity (Tox), AD and silencing (SD) domains [125]. Among these, both the DBD and the C-terminal portions of Rap1 (containing Tox, AD and SD domains) interact directly with the holo-TFIID complex, indicating that the C-terminal domain of Rap1 likely contributes to the observed external density [42]. In a control experiment we examined the structure of isolated Rap1 molecules and found that they adopt a two-lobed horseshoe shaped structure, consistent in size and shape with the two densities found in lobe B. We can however not completely rule out the possibility that this second density corresponds to a second Rap1 molecule. In addition to Rap1, Complex I contains an extra density associated with TFIID domain D near the Taf2 site (Lobe D, green in **Fig. 8a**). A similar density was detected at the same location in the TFIID-TFIIA-DNA complex suggesting that this density corresponds to DNA bound to Taf2. Thus, in Complex I, TFIID appears to contact DNA through Rap1 and Taf2, but apparently not through TBP and TFIIA, since no extra density was detected in lobe C1. Taken together these data suggest that complex I may correspond to an initial "recruitment mode" of binding, where TFIID interacts with the promoter-bound activator in the absence of TFIIA.

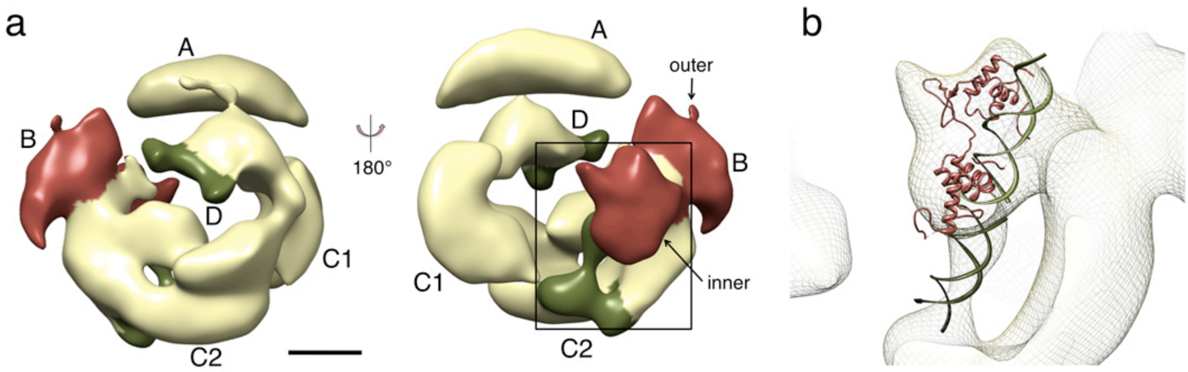


Figure 8 Structure of the initial TFIID-Activator-Promoter recruitment complex.
a, Two different surface views of the Cryo-EM map of Complex I formed upon incubating TFIID, TFIIA, Rap1 and the *UAS$_{RAP1}$-PGK1* enhancer-promoter DNA. TFIIA is not detected in Complex I. Densities originating from Rap1 are detected on both sides of lobe B and are colored red. Densities attributed to DNA are colored green in lobes D and C2. **b**, Enlargement of the area boxed in (a) and fitting of the atomic model of DNA-bound Rap1 DBD into the additional Rap1-density contacting the inner face of lobe B. The rod of additional density protruding towards lobe C2 superimposes to the expected position of Rap1-bound DNA.

Complex II, a quaternary complex containing TFIID, Rap1, TFIIA and enhancer promoter DNA, is characterized by a continuous density between TBP and the Rap1 binding sites that bridges lobes C1 and B (**Fig. 9a**). This bridge is too large to be composed of DNA alone and therefore must also contain protein. We propose that the mass next to lobe C1 corresponds to TFIIA, as TFIIA binds to a similar site of TFIID in the TFIID-TFIIA-DNA complex, and because this extra mass accommodates the atomic structure of the TFIIA-TBP-DNA complex [126] (**Fig. 9b**). Notably though, comparison of Complex II with the TFIID-TFIIA-DNA complex, indicates that the position of TFIIA is rotated 130° around its TFIID interaction site (**Fig. 9c**). The repositioning of TFIIA within Complex II brings it close enough to directly interact with the DBD of Rap1, which is located on the inner face of lobe B as in Complex I. The proposed position of the Rap1 DBD is supported by the docking of its atomic structure into the remaining part of the protein bridge (**Fig. 9b**). Interactions between activators and TFIIA have been described [127-130] and in one case TFIIA was shown to be required to release TBP from DNA binding autoinhibition mediated by the N-terminal TAND domain of Taf1 [131], whereas in the other instances TFIIA was described to stimulate activator dependent transcription by interacting with the activator, the situation we observe here in Complex II. This reorganization of TFIIA within the TFIID-activator-promoter complex likely affects the position and therefore the accessibility (and functionality) of TBP, contributing to both increased TFIID-promoter interaction, PIC formation, and ultimately initiation efficiency. The exact path of promoter DNA cannot be traced in Complex II, though several extra densities signal its position. A patch of density is found on the inner wall of the clamp between lobes C1 and D, and a second more robust density is located in lobe D as observed in both Complex I and the TFIID-TFIIA-DNA complex (Lobe D, green in **Fig. 9a**). We interpret these stretches of density to correspond to the bound promoter DNA between the TATA box that interacts with TBP and the DNA close to the transcription start site bound by Taf2.

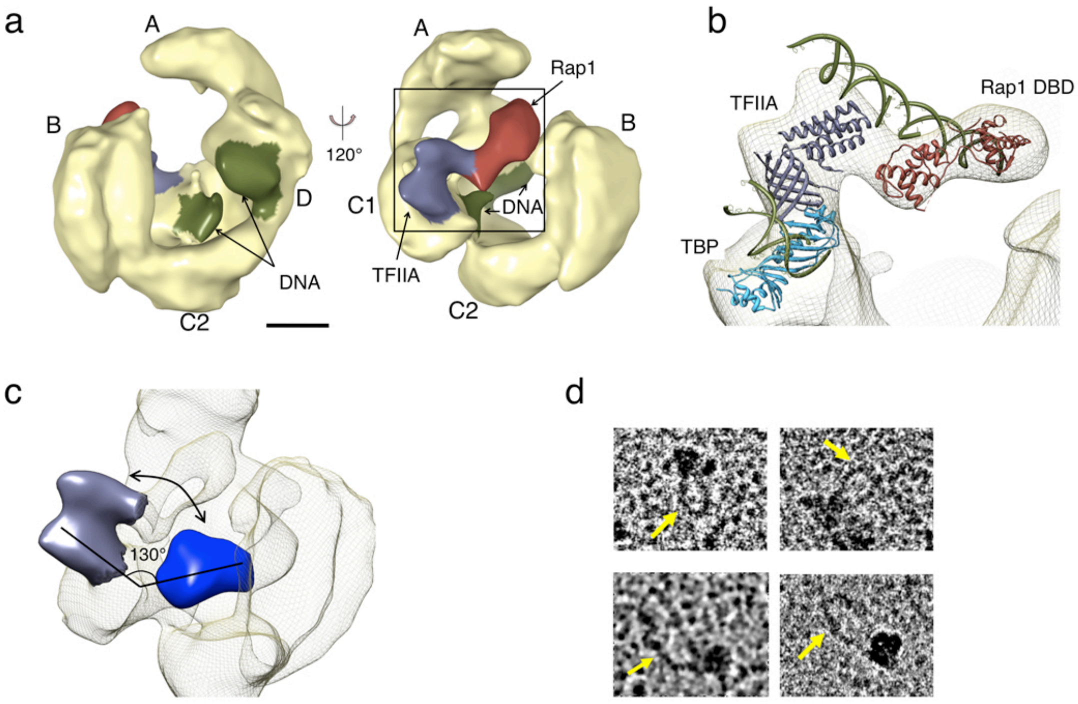


Figure 9 Structure of the committed complex.

a, Two different surface views of the Cryo-EM map of Complex II formed upon incubating TFIID, TFIIA, Rap1 and *UAS*$_{RAP1}$-*PGK1* enhancer-promoter DNA. The additional densities revealed in Complex II are colored as follows: DNA, TFIIA and Rap1 are depicted in green, violet and red, respectively. **b**, Enlargement with slight tilting of the area boxed in (a) and fitting of the crystal structure of the TBP-TFIIA and the Rap1-DBD-DNA complexes identifies the bridging density between lobes C1 and B. Note that part of TFIIA is missing in the crystal structure and may affect the fitting **c**, Comparison of the position of TFIIA between the TFIID-TFIIA-DNA complex (blue) and Complex II (violet) reveals that the position of TFIIA is rotated by 130°. **d**, Platinum shadowing of spread TFIID-TFIIA-Rap1-DNA complexes showing the formation of a DNA loop in the presence of Rap1.

The DNA between the TATA-box and the Rap1 binding sites was not detected, most likely because of the flexibility of this segment. The proposed arrangement of the Rap1 DBD and TBP/TFIIA complexes implies that the intervening DNA loops out away from TFIID. To test this hypothesis we formed complexes on the yeast enhancer-promoter DNA fragment with TFIID and TFIIA in the presence and absence of Rap1. The resulting complexes were visualized after platinum shadowing and clearly revealed loops of DNA protruding from TFIID (**Fig. 9d**). Rap1 plays an essential role in loop formation since in the absence of Rap1, loops were observed in less than 2% of the DNA-bound TFIID complexes, while when Rap1 was added, 35% of the DNA bound TFIID molecules had DNA loops. Similar Rap1-dependent loop formation is observed on the natural *RPS1A* gene. Moreover TFIIA mutants unaffected in TBP binding but showing impaired transcription show a reduced ability to form DNA loops. These observations demonstrate that the Rap1 activator favours the formation of DNA loops allowing communication with distant TFIID-bound promoter sequences.

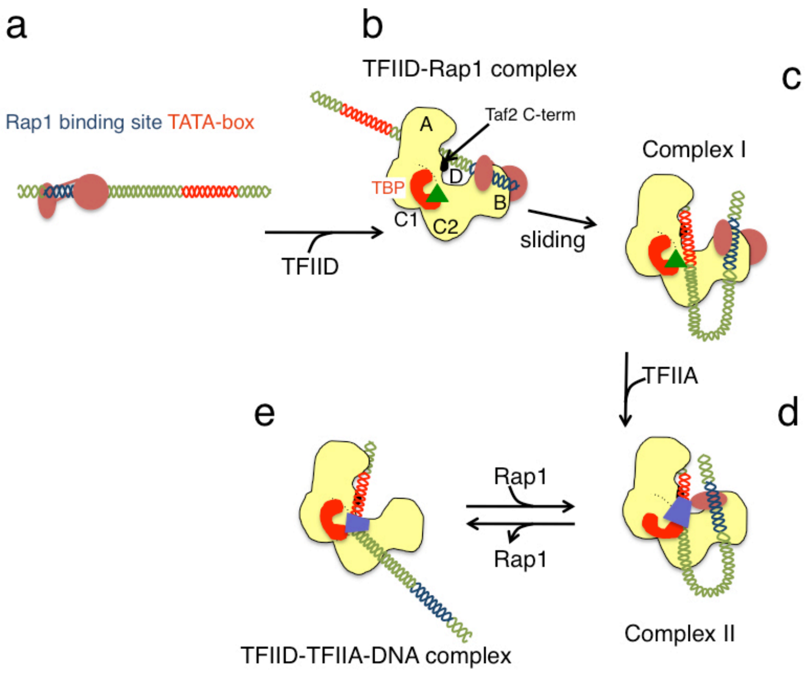


Figure 10 Model depicting the formation of the activated TFIID complex.
a, Binding of Rap1 (red circles) to its specific DNA recognition elements. **b**, Recruitment of TFIID (yellow) through an interaction with Rap1 and Taf2 (black dot). **c**, Formation of a DNA loop. **d**, Recruitment of TFIIA (blue trapezoid) which induces the formation of a protein bridge between lobes B and C1 that locks the DNA loop. **e**, Model showing the different position of TFIIA in the TFIID-TFIIA-Ad2 MLP DNA complex, which naturally lacks Rap1.
The red shape corresponds to TBP while the green triangle represents TAND autoinhibition.

We propose that Complexes I and II represent functional intermediates in the pathway leading to PIC formation. From the molecular snapshots we have captured we infer a possible mechanism for activated

TFIID-TFIIA-DNA complex formation (**Fig. 10**). In a first step, Rap1 interacts with its recognition element (**Fig. 10a**) and this Rap1-enhancer complex can simultaneously bind TFIID (**Fig. 10b**). In all complexes analyzed, the Taf2–containing D lobe was found to interact with the DNA template indicating that in the absence of detectable interaction with TBP or TFIIA, the DNA is already looped out (**Fig. 10c**). Whereas the role of TFIIA in releasing TBP autoinhibition in basal transcription is well established, its contribution to activated transcription at the molecular level is less well understood. The discovery of a class of TFIIA mutants that stimulate TBP-DNA binding but fail to support activation favours a model in which TFIIA acts in two mechanistically distinct activation steps [129] consistent with the structures reported here (**Fig. 10c,d,e**), where we propose that the action of the activator is mediated by a protein bridge between lobes B and C1 of TFIID through an interaction between TFIIA and the Rap1 DBD (**Fig. 10d**). Thus activator-induced repositioning of TFIIA, and probably of its interaction partner TBP, may affect the accessibility of the DNA binding surface of TBP thereby facilitating functional PIC formation and activation of transcription (Complex II, **Fig. 10d**). In addition our model predicts that the bridge closes over the TBP-bound DNA and topologically locks the proximal promoter DNA in the resulting clamp (Complex I to Complex II; **Fig. 10c,d**). Such a trapping process could result in an increased residence time of the promoter DNA within TFIID and participate in the activation mechanism. Collectively our results support a role for TFIID as an assembly platform that plays an active and important role in PIC formation and transcription. Our data provide new structural insights into how, Rap1 collaborating with TFIIA, transduces activating intramolecular signals within the TFIID coactivator complex that ultimately can lead to PIC formation. Further, the structures reported here highlight the complex network of protein-protein and protein-DNA interactions regulating activated transcription.

Methods

Protein purification and DNA probes: HA$_1$-Taf1- and TAP-Taf1-tagged TFIID were purified from *Saccharomyces cerevisiae* as described previously [1, 4, 95]. Rap1 and TFIIA were expressed in *E. coli* and purified as described[4, 42, 120]. Two promoter-containing DNA fragments were used to form complexes. The TFIID-TFIIA-DNA complex was assembled on a 400 bp long fragment of the Ad2 MLP, while a 282 bp long *UAS$_{RAP1}$-PGK1* chimeric yeast promoter was used for the assembly of the TFIID-TFIIA-Rap1-DNA complex. This promoter fragment contains the 41 bp *UAS$_{RAP1}$* element, derived from the ribosomal protein-encoding *RPS8A* gene (from -252 to -212), which contains two binding sites for Rap1 (CTT**TACATCCATACACC**CTCTTT**AACACCCTTACACT**TTTA; Rap binding sites bold, underlined) fused to the *PGK1* core promoter (-211 to +30)[42, 122, 132].

Sample preparation and electron microscopy: The final concentration of TFIID used in the negative stain and cryo-EM experiments was 30 µg/ml and of 50 µg/ml respectively. The crosslinked samples were treated with 0.1% glutaraldehyde for 5 sec prior to adsorption on a thin carbon film. To assemble the TFIID-Rap1 complex Tap-tagged TFIID was incubated with a 5-fold molar excess of Rap1 for 30 min on ice in 20 mM HEPES pH 7.9, 250 mM NaCl, 1 mM DTT, 0.2 mM PMSF and 10% glycerol. The sample was adsorbed onto air glow-discharged grids covered with a 10 nm thick carbon film and sandwiched with a second carbon film after negative staining with 2% uranyl acetate. A transmission electron microscope (TEM, Philips CM120) equipped with a LaB$_6$ cathode and operating at 120 kV was used to collect images at 45,000X magnification on a Pelletier-cooled slow scan CCD camera (Model 794, Gatan, Pleasanton, CA) resulting in a pixel spacing of 0.37 nm on the object. To assemble the TFIID-TFIIA-DNA complex, TAP-tagged TFIID was incubated for 20 min at 20°C with a 40-fold molar excess of TFIIA and 4-fold molar excess of the Ad2 MLP in 10 mM HEPES pH 7.4, 60 mM KCl, 6 mM $MgCl_2$ and 2 mM DTT. The specimen was adsorbed on a holey carbon grid covered with a 3-4 nm thick carbon film. Images were recorded on a cryo-TEM (FEI Tecnai F20) equipped with a field emission gun (FEG) and operating at 200 kV. Images of well dispersed individual complexes were recorded at liquid nitrogen temperature on Kodak SO-163 films at 40,000 X magnification and in low dose conditions (15-20 e-/$Å^2$). Negatives were digitized with a 5 µm raster size using a drum scanner (Primescan D7100, Heidelberg) and were coarsened twice to obtain a pixel spacing of 0.254 nm on the object. For the TFIID-TFIIA-Rap1-DNA complex, HA-tagged TFIID was incubated 30 min at 20 °C with a 10-fold molar excess of *UAS_{RAP1}-PGK1* enhancer-promoter DNA, a 5-fold excess of TFIIA and a 5-fold excess of Rap1 in 10 mM Tris-HCl pH 7.9, 170 mM KOAc and 5 mM $MgCl_2$. The specimen prepared as detailed above for the TFIID-TFIIA-DNA complex, was vitrified in Vitrobot (FEI) and observed with a cryo-TEM (FEI Tecnai Polara) equipped with a FEG (field emission gun) operating at 300 kV. Images were collected at liquid nitrogen temperature under low-dose condition (15-20 e/$Å^2$), at a magnification of 39,000 X on Kodak SO-163 films. The pixel spacing of the digitized negatives was of 0.26 nm.

DNA loop formation was visualized after absorption of the DNA-protein complexes onto air glow-discharged grids that were positively stained with uranyl-acetate and platinum shadowed after air-drying.

Image processing: Boxing of the images of the TFIID-Rap1 and TFIID-TFIIA-DNA complexes was performed with the EMAN software [133] package whereas the images of TFIID-TFIIA-Rap1-DNA complex were selected with the boxer2 option of EMAN2 [134]. The contrast transfer function (CTF) of the microscope was estimated using Bsoft:Bshow [135] and the images were corrected by phase flipping. Image processing was performed using the IMAGIC [112] (Image Science Software, Berlin, Germany) and Spider [113] software packages as described earlier [1]. The resolutions of the final reconstructions were estimated according to the intersection point of the half bit curve with the FSC curve (half bit criterion)[114]. The final reconstructions were filtered to the estimated resolution.

In order to improve specimen homogeneity, TFIID-containing complexes that did not contain Taf2 were removed from the dataset [1]. During refinement, images of the TFIID-Rap1 and TFIID-TFIIA-DNA complex were split into holo-TFIID and complex. To do so, images were sorted for their best cross-correlation with reprojections of either the mixed (holo-TFIID + complex) model or the holo-TFIID reference model. This separation was iterated several times and resulted in a progressive enrichment of the mixed model in TFIID-Rap1 or TFIID-TFIIA-DNA complexes. The analysis of TFIID-TFIIA-Rap1-DNA complex was similar except for a 3-D statistical analysis and clustering step that was performed to separate distinct conformational states according to ref 24. Briefly, a large number of 3-D models were reconstructed from a few randomly selected and pre-aligned class average images. This repertoire of 3-D models was subjected to multivariate statistical analysis and was clustered into groups corresponding to different conformations of the complex. The class-sum volumes, characteristic for each conformation, were used as references for subsequent refinement rounds. Fitting of atomic coordinates into EM density maps were performed using UCSF Chimera's fit in map tool [111] and Sculptor [136]. Images were created with UCSF Chimera.

THE ARCHITECTURE OF HUMAN GENERAL TRANSCRIPTION FACTOR TFIID CORE COMPLEX

Results

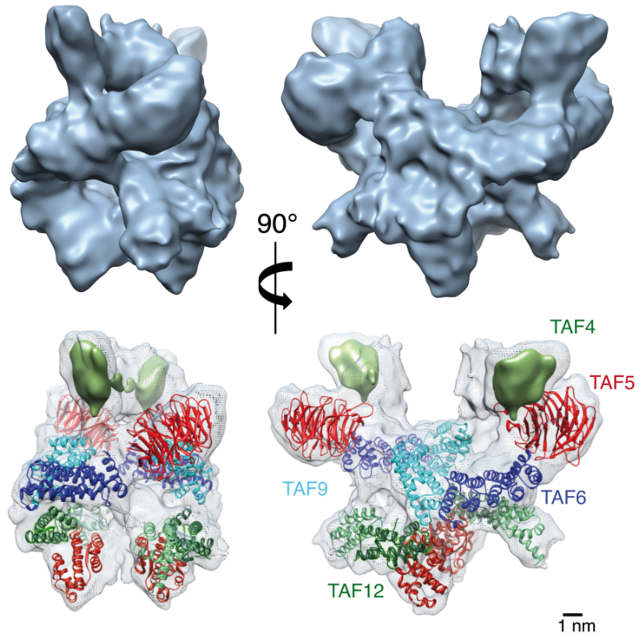


Figure 11 Structure of the human TFIID core complex.
The cryo-EM structure (top) is displayed in a side view (left) and from the front (right). The structural features in core-TFIID are shown (bottom). The cryo-EM density is transparent, TAF5 is colored red (WD40 repeat domain, N-terminal domain), the TAF6 C-terminal domain is dark blue, the TAF6 and TAF12 HF pair is light blue. TAF4 N-terminal part, TAFH-domain and HF pair with TAF12 are colored green.

The overall shape of human and yeast TFIID was unveiled by electron microscopy, revealing an asymmetric tri-lobed structure resembling a molecular clamp [93, 136]. The paucity and heterogeneity of the endogenous material used in these studies limit structural insight to moderate resolution (~30Å for human TFIID), prohibiting molecular level interpretation of TFIID architecture [2, 137]. Endogenous yeast TFIID was analyzed for subunit stoichiometry, revealing that a subset of six TAFs (TAF4, 5, 6, 9, 10 and 12) exist in two copies, while TBP and the remaining seven TAFs are present in single copy [4]. A similar subunit composition of the human complex is likely, owing to conservation and overall resemblance in shape of yeast and human TFIID [137]. The concept emerged in which TAFs present in duplicate form a two-fold symmetric scaffold, around which the remaining TAFs and TBP organize as peripheral subunits [107, 136]. Compelling functional support came from studies in *Drosophila* cells, revealing a functional core-TFIID complex, composed of TAF4, 5, 6, 9 and 12 *in vivo* [107]. In cryo-electron microscopy (cryo-EM) studies of yeast TFIID preparations, a quasi-symmetric smaller shape was also found [1]. Together, these results suggest the existence of a core-TFIID module of pivotal importance for the integrity and assembly of holo-TFIID [136].

We expressed and purified recombinant human core-TFIID complex, consisting of two copies each of TAF4, 5, 6, 9 and 12. We determined the structure of this ~650 kDa complex by single-particle cryo-EM (**Figure 11**). The presence of two copies each of the five TAFs is suggestive of a symmetric core-TFIID architecture. We scrutinized this by carrying out a complete refinement without applying any symmetry constraint. The resulting structure exhibits two-fold symmetry at 13.4 Å resolution. We refined the structure by imposing this symmetry constraint and reached a resolution of 11.6 Å.

The structure of human core-TFIID complex reveals an interlaced architecture and a remarkably large solvent accessible surface due to numerous protrusions and channels (**Figure 11**). An iterative density truncation approach allowed us to place all conserved domains of the TAFs within core-TFIID. By fitting coordinates from crystal structures or homology models, and by biochemical engineering of key subunits, we could assign ~70% of the density to specific TAF domains.

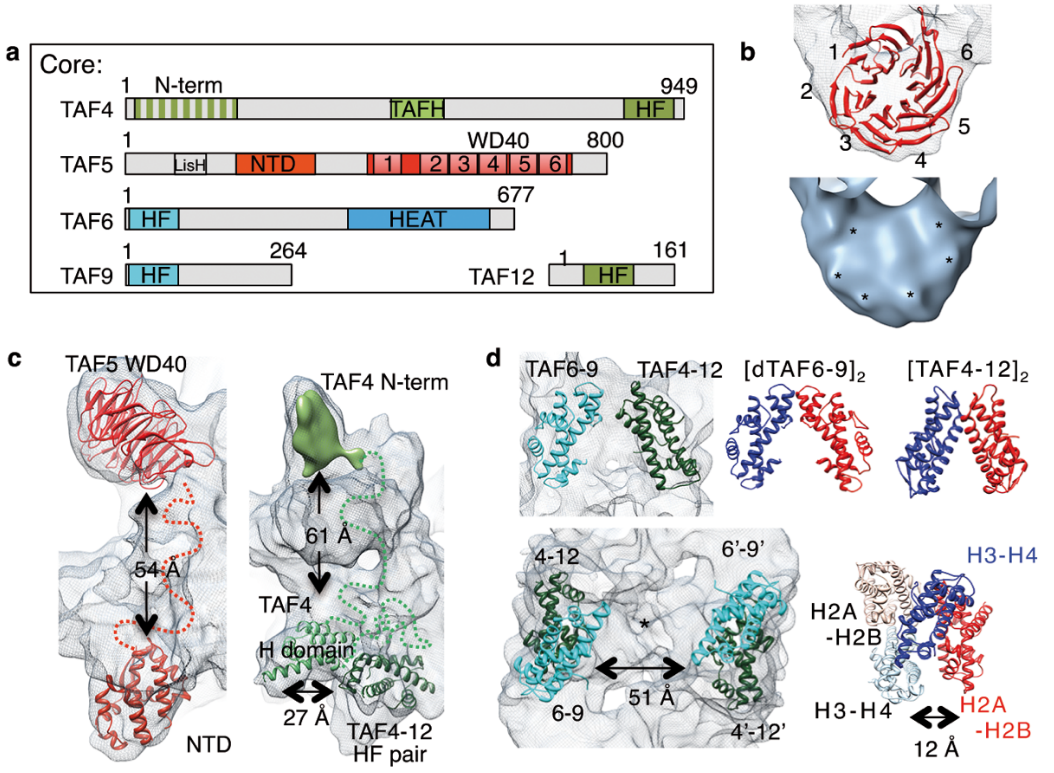


Figure 12 Molecular organization of conserved TAF domains.
a, TAF domain architecture (color code as in Figure 1). NTD stands for TAF5 N-terminal domain, WD40 for TAF5 C-terminal WD40 repeats. The TAF4 N-terminal part is hatched. HF: histone fold; HEAT: TAF6 C-terminal HEAT repeats; TAFH: conserved peptide-interaction domain in TAF4. LisH: non-conserved homology region. **b**, Density corresponding to the TAF5 WD40 repeat domain (top), with a closely related β-propeller (PDB entry 2PBI) superimposed (red). Six knuckles are marked by asterisks (bottom). **c**, Conformations adopted by TAF5 (red) and TAF4 (green) in core-TFIID. Amino acid stretches with unknown conformation are represented as dotted lines. **d**, Structural arrangement of TAF6/TAF9 and TAF4/TAF12 HF pairs, looking at the front (top, left). An identical arrangement is in the back. TAF6/TAF9 and TAF4/TAF12 tetramers present in crystals (top, right) and the histone octamer (bottom, right) are depicted for comparison. The distance relating H2A/H2B and H3/H4 (12 Å) and the pseudo two-fold axis in the octamer (dashed line) are indicated. The HF pairs at the front and back of the core-TFIID structure are separated by >50 Å (bottom, left). The core-TFIID two-fold axis is marked (asterisk).

In the core-TFIID cryo-EM structure, a flat, slightly conical shape projecting from either side exhibits clear features of β-propeller structures characteristic of WD40 repeat domains. The TAF5 C-terminal region contains six predicted WD40 repeats (**Figure 12a**) [50]. Our density shows six triangular knuckles consistent with six blades (**Figure 12b**). TAF5 also contains a conserved N-terminal domain (NTD) for which crystal structures exist [138, 139]. Two protrusions at the bottom of the core-TFIID structure accommodate the crystal coordinates, suggesting that the TAF5 NTD is located distally from the C-terminal domain comprising the WD40 repeats (**Figure 12c**). It has been hypothesized that the TAF5 NTD may play a role in dimerization of TAF5 [138]. In our structure, the TAF5 NTDs are not sufficiently close to engage an extended dimerization interface.

A recent crystal structure revealed a HEAT repeat domain in the TAF6 C-terminal part, consisting of ten tightly packed α-helices [140]. The TAF6 HEAT repeats are located adjacent to the TAF5 WD40 repeat domains, bracketing the front and back of the complex (**Figure 11**). TAF6 also contains a conserved histone fold (HF) domain, which specifically interacts with a HF domain present in TAF9 to form a HF pair (**Figure 12a**) [56]. The cryo-EM density of core-TFIID exhibits four regions that can accomodate HF pairs, two TAF6/TAF9 and two TAF4/TAF12 heterodimers. In order to assign their location, we determined the structure of a previously characterized ~400 kDa heterohexameric complex [141] containing two copies each of TAF5, 6 and 9 (denoted hereafter 3TAF) by cryo-EM and single-particle analysis. The 3TAF density reveals a holey basket-like structure with dimensions similar to core-TFIID, however lacking protrusions. The TAF5 WD40 repeat and NTD domains as well as the TAF6 HEAT repeats, and also the TAF6/TAF9 HF pairs are clearly discernible in the 3TAF structure, enabling unambiguous assignment of the TAF6/TAF9 HF pairs in core-TFIID. Note that we placed the pair as a unit as we cannot discriminate the TAF6 HF from the TAF9 HF. We also determined the cryo-EM structure of a mutant 3TAF complex containing TAF5 N-terminally tagged with maltose binding protein (MBP) to confirm the TAF5 NTD placement.

TAF4 contains a N-terminal region of apparent low complexity, a central conserved domain called TAFH, and a conserved HF domain in the C-terminal region (**Figure 12a**). The TAF4 HF domain pairs with TAF12 and atomic structures have been determined [58]. The difference density map between core-TFIID and 3TAF readily allowed us to assign the position of the two TAF4/TAF12 HF pairs, occupying density adjacent to the TAF6/TAF9 HF pairs. TAFH binds short hydrophobic peptides present in transcriptional regulators and the crystal structure shows a compact bundle of α-helices [142]. Two protrusions in the neighborhood of the TAF4/TAF12 HF pairs accommodate the TAFH crystal coordinates (**Figure 12c**). We further dissected the structure of TAF4 by engineering a mutant core-TFIID containing N-terminally truncated TAF4, and we determined the EM structure of this complex. The two ear-like protrusions straddling the TAF5 WD40 repeat domains disappeared, indicating that this density corresponds to the TAF4 N-terminal parts. In contrast, the lateral protrusions in the lower part of the complex remained unaltered, confirming our TAFH placement.

The discovery of histone fold motifs in TAFs spawned intense discussion on the contribution of HFs to the integrity of TFIID. The crystal structure of the *Drosophila* TAF6/TAF9 HF pair showed structural similarity with the heterotetrameric core of the histone octamer, formed by histones H3 and H4 (**Figure 12d**) [56]. Biochemical data suggested a similarity of TAF4 and TAF12 to histones H2A and H2B, respectively, leading to the proposal that a histone octamer-like structure may exist within TFIID [55, 143]. Our cryo-EM structure of human core-TFIID contains two copies each of the TAF6/TAF9 and TAF4/TAF12 HF pairs. In the front and in the back of core-TFIID, one TAF6/TAF9 pair is juxtaposed to one TAF4/TAF12 pair, reminiscent of the tetramers in the crystal structures, the latter however consist of two identical dimers (**Figure 12d**). In contrast, the TAF6/TAF9 and TAF4/TAF12 HF pairs are less tightly associated and rotated with respect to each other. The distance across core-TFIID (>50 Å) rules out direct interactions between the two sets of HF pairs, whereas in the histone octamer, the H2A/H2B and H3/H4 pairs are within van-der-Waals contact. Our results suggest that histone octamer-like arrangements mediated by TAF4, 6, 9 and 12 are not formed in TFIID, while underscoring the important architectural role of the HF as a strong protein-protein interaction motif.

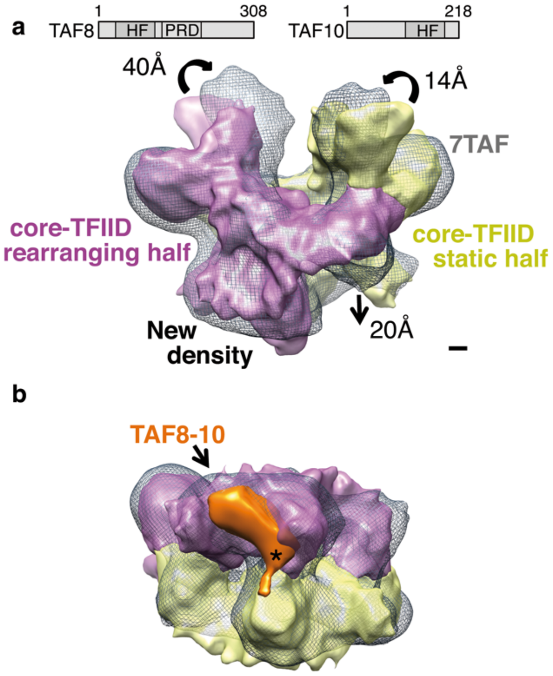


Figure 13 Asymmetric 7TAF structure. **a**, TAF8 and TAF10 are represented as bars (top) showing HFs and a proline-rich domains (PRD). The 7TAF structure (grey mesh) is superimposed on core-TFIID colored in purple (rearranging half) and yellow (static half). Conformational changes are marked with arrows. New density is observed in 7TAF. Scale bar: 1nm **b**, 7TAF (mesh) in a bottom view, superimposed on core-TFIID. New density in the 7TAF structure is drawn in orange. The two-fold axis of core-TFIID is marked (asterisk).

Previous analyses of TAF locations relied on antibody mapping of yeast endogenous TFIID [3, 91]. Our TAF5 geometry in core-TFIID is consistent with the immuno-mapping data, which detected two copies of TAF5, placed their N-terminal regions in close proximity, and mapped the C-terminal domains with the WD40 repeats to two opposite lobes in the holo-complex [3]. Likewise, the immuno-mapping study identified two copies of TAF6 and TAF9, one each in the vicinity of the C-terminal parts of TAF5. On the other hand, the TAF6-TAF9 and TAF4-TAF12 pairs arrange symmetrically in core-TFIID, while the immuno-mapping studies suggested an asymmetric arrangement [3]. This discrepancy may reflect errors in the immuno-mapping experiments. Alternatively, the differences may stem from changes in conformation or accessibility of the core-TFIID subunits, when further TAFs are accreted and holo-TFIID is formed. Further, the immuno-mapping experiments utilized sample from yeast, while the present core-TFIID structure is from human. Yeast and human TAFs exhibit considerable variation in size, which may also affect their geometries.

The structure of core-TFIID contains two copies each of its subunits in a symmetrical arrangement, while holo-TFIID, which contains additional TAFs and TBP, has the shape of an asymmetric clamp [137]. How does the structural transition occur from a symmetric to an asymmetric state during TFIID assembly? TAF8 regulates the nuclear import of TAF10, and both TAFs were shown to be co-imported as a complex into the nucleus by an importin α/β dependent pathway [144]. Combinatorial assembly experiments showed that the TAF8/TAF10 pair can only be incorporated into a larger complex when all five TAFs forming core-TFIID are present [145]. Thus, we hypothesized that the transition from a symmetric core-TFIID to an asymmetric assembly may occur at the step of TAF8/TAF10 complex integration and may be regulated by nuclear import mechanisms. We prepared a complex comprising core-TFIID and TAF8/TAF10, and determined the structure of this ~710 kDa complex (hereafter denoted 7TAF) by cryo-EM (**Figure 13**).

The 7TAF complex structure shows major perturbations when compared to core-TFIID and notably deviates from two-fold symmetry (**Figure 13**). Careful inspection of the cryo-EM density reveals that two different parts can be defined in the 7TAF structure. One half of the 7TAF complex adopts largely the same shape as in core-TFIID. In contrast, most of the structural rearrangements localize to the other half (**Figure 13a**). The ear-like lobe on top, swings over by ~40Å in this rearranged half. In addition, new density was observed at the bottom in the vicinity of the TAF5 NTD. We reasoned that this new density could be attributed to the binding of TAF8/TAF10. We used a mouse monoclonal antibody (mAb6TA) that specifically binds TAF10, to prepare a 7TAF-mAb6TA complex. EM analysis revealed binding of the antibody to the bottom part of the structure, confirming the position of the TAF8/TAF10 complex. The new density is consistent with the volume occupied by one HF pair suggesting that a single TAF8/TAF10 heterodimer is incorporated into the 7TAF complex. Reconstitution experiments of core-TFIID and wild-type TAF8/TAF10 complex at defined ratios were consistent with the presence of one copy of TAF8/TAF10 in the 7TAF complex, and experiments involving a MBP-tagged variant of TAF8 confirmed this finding. The new density in the 7TAF complex extends over the two-fold axis that previously related the two halves of core-TFIID (**Figure 13b**). Steric hindrance thus rules out the incorporation of a second TAF8/TAF10 copy. We demonstrated that the stoichiometry of TAFs in our recombinant complexes, notably the existence of a single copy of TAF8, is the same as in endogenous human TFIID by comparative Western blots, and by protein abundance determination following mass spectrometry analyses.

Our results provide a mechanistic model for the structural transition of TFIID from a symmetric core to the asymmetric holo-complex. TAF8 and TAF10, co-imported as a complex into the nucleus, bind to symmetric core-TFIID, and once integrated, prevent recruitment of a second TAF8/TAF10 pair, thus breaking the symmetry in core-TFIID and inducing major conformational changes. The presence of TAF8/TAF10, combined with the rearrangements, give rise to two structurally distinct halves and create new surfaces for the incorporation of the remaining TAFs and TBP in single copy, to complete the asymmetric holo-TFIID (**Figure 14**).

Our core-TFIID structure identifies, as a central scaffold, two copies each of TAF5, 6 and 9, held together by multiple interactions involving the HFs, the TAF5 NTD and direct interfaces between the TAF5 WD40 repeat domain and the TAF6 C-terminal part. The TAF6/TAF9 pair has been shown to bind the downstream core promoter element (DPE) in TATA-less promoters specifically [47]. Our structure places the TAF6/TAF9 pairs to the surface, well positioned to interact with DPEs. TAF5, 6 and 9 were identified in both TFIID and SAGA in yeast. SAGA is a further important large coactivator multiprotein complex found at promoters [146]. In humans, gene duplication resulted in TAF5L and TAF6L, which are closely related variants substituting for TAF5 and TAF6 in human SAGA [146]. We propose that the 3TAF structure constitutes the common central scaffold of TFIID and SAGA, around which the other subunits assemble in both complexes.

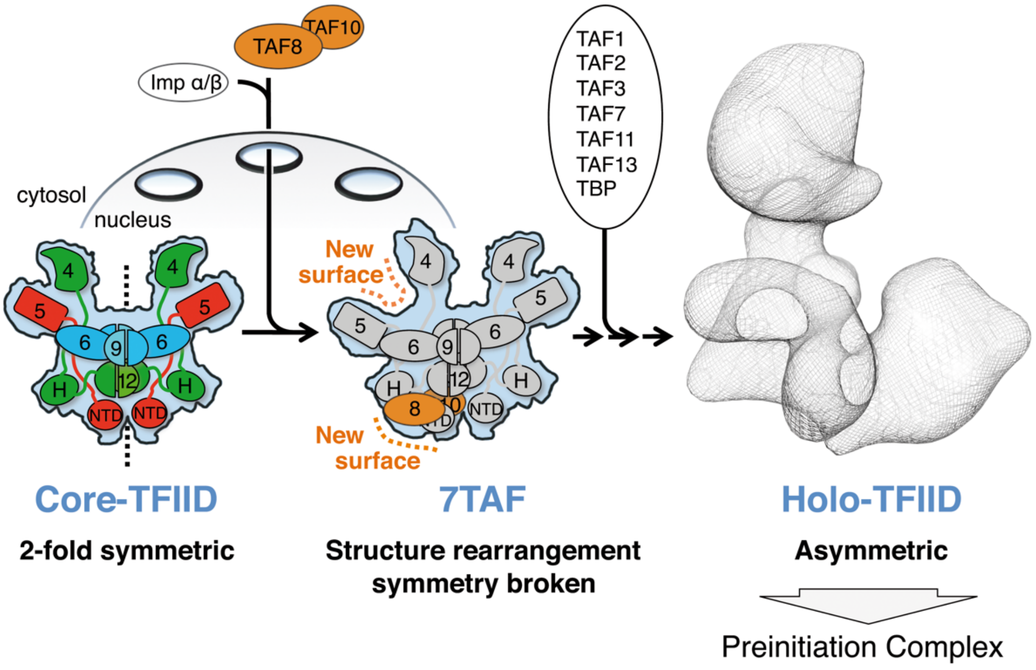


Figure 14. Model for holo-TFIID assembly.
Core-TFIID with two copies of TAF4, 5, 6, 9 and 12, is symmetric (left). TAF8/TAF10 complex (orange) is imported into the nucleus by importins[23] (top). Binding of one copy of TAF8/TAF10 breaks the symmetry in core-TFIID, resulting in asymmetric 7TAF complex (middle). 7TAF exhibits two distinct halves and new binding surfaces for further subunits (dashed lines). Accretion of remaining TAFs and TBP in single copy, results in asymmetric clamp-shaped holo-TFIID (EMD-1195, grey mesh) that nucleates the PIC (right).

TAF variants mediating specific cellular functions have been identified [102, 147-149]. Incorporation of the cell-type specific TAF4b was shown to induce an open conformation in TFIID, potentially facilitating activator binding [102]. In cells where TAF4b is expressed, TAF4 and TAF4b co-exist in TFIID complexes. As TAF4b contains a different N-terminal region than TAF4, the structural basis of the open conformation of TAF4b-containing TFIID complexes may reside in the distinct geometries of the ear-like lobes. TAF6δ was found to link apoptotic signaling pathways to TFIID function [148]. TAF6δ has a deletion in its HF domain, and TAF6δ-containing TFIID was suggested to lack the HF partner of TAF6, TAF9 [148]. Our 3TAF structure shows numerous interfaces between two copies each of TAF5, 6 and 9, and the loss of TAF9 due to the compromised HF in TAF6δ may be tolerated to some extent during TFIID assembly. The TAF9 related factor, TAF9b, was implicated in gene silencing and transcriptional repression [149]. TAF9b and TAF9 have very similar sequences, and we expect that incorporation of TAF9b will not cause major rearrangements, implying reasons other than structural for its specific functions.

We determined the structures of three distinct TAF subcomplexes of human TFIID, providing a molecular framework for rationalizing TFIID core architecture. Our results support the concept of a unique and step-wise assembly pathway for holo-TFIID and functional intermediate TFIID complexes

[34, 107, 136, 145], which may have specific roles in gene transcription [34]. Conformational changes are found between the structures, most pronounced when 7TAF is formed from core-TFIID and the TAF8/TAF10 complex, breaking the original symmetry in the process. These structural changes likely recapitulate molecular events along an assembly pathway, when TFIID is crafted in the cell.

Methods

Recombinant TAF complexes were produced utilizing MultiBac and polyproteins. Endogenous TFIID for mass spectrometry was purified from cultured HeLa or fetal liver cells. Negative-stain EM was performed and 2D class averages were calculated (IMAGIC software package) as a benchmark for optimizing complex purification protocols until satisfactory sample quality was achieved. For cryo-EM grid preparation, TAF complexes were stabilized by mild glutaraldehyde cross-linking in a glycerol gradient. The specimens were adsorbed on a thin carbon film sustained by a holey-carbon grid and plunge-frozen in liquid ethane with controlled temperature and humidity. Images of the 3TAF and core-TFIID complexes were recorded at 50,000x on a cryo-transmission electron microscope (cryo-TEM, Tecnai F20, FEI) at 200 kV, digitized on a drum scanner (Primescan D7100, Heidelberg) at 5000 dpi. 3TAF was coarsened twice resulting in a pixel spacing of 2.03 Å, core-TFIID was coarsened three times resulting in a pixel spacing of 3.05 Å before analysis on the specimen. CCD frames of the 7TAF complex were recorded at 58,000x on a cryo-TEM (Tecnai Polara, FEI) operating at 100 kV and coarsened twice resulting in a final pixel spacing of 3.72 Å. Images were selected using the EMAN2 software package and analyzed with the IMAGIC, Bsoft and Spider software packages. The resolution of the structures was determined according to the 0.5 cutoff of the Fourier Shell Correlation curve and the final reconstructions were filtered accordingly. Interactive fitting of atomic structures and generation of images for publication were performed using the UCSF Chimera software.

MOLECULAR STRUCTURE OF PROMOTER-BOUND YEAST TFIID

Here we describe the cryo-EM structure of the yeast *Komagataella phaffii* (formerly known as *Pichia pastoris*) TFIID complex at a resolution of 4.5 to 10.7 Å. We position the crystal structures of several subunits and, in combination with cross-linking studies, describe the quaternary organization of TFIID. The compact tri lobed *Kp*TFIID architecture is stabilized by a topologically closed $(Taf5\text{-}Taf6)_2$ tetramer. We confirm the unique subunit stoichiometry prevailing in TFIID and uncover a hexameric arrangement of Tafs containing a HF domain. Interaction with promoter DNA highlights two non-selective binding sites consistent with a DNA scanning mode.

Results

Currently available structural information on the complete yeast TFIID complex is derived from low-resolution electron microscopy studies precluding docking of existing atomic models [3, 25]. We developed a TFIID purification protocol from the yeast *K. phaffii* for structure determination by high-resolution cryo-EM. SDS-PAGE experiments (**Fig. 15a**) and mass spectrometry analysis (data not shown) confirm that the TFIID subunit composition of the phylogenetically closely related *S. cerevisiae* and *K. phaffii* are identical. To gain a better understanding of the subunit organization within TFIID, a CXMS analysis was performed using the homo-bifunctional, amine-reactive crosslinking reagent bis(sulfosuccinimidyl)suberate (BS3). A total of 391 unique intra-protein crosslinks (intralinks) and 488 unique inter-protein crosslinks (interlinks) were identified. The dense BS3 crosslinking map underlined many known protein-protein and domain-domain interactions within the TFIID complex (**Fig. 15b and c**). For example, Taf7 is extensively crosslinked to Taf1, in agreement with previous studies showing that these two subunits directly interact [150]. Genetic and biochemical evidence indicated that Taf5, Taf4, Taf12, Taf6 and Taf9 are present as two copies in TFIID [4] and a recombinant human core TFIID containing these 5 Tafs shows two-fold symmetry [26]. Taf5 crosslinks extensively to all the core subunits (**Fig. 15c**), suggesting that this subunit serves as an organization center for core subunits assembly. Nine Tafs contain a protein fold homologous to nucleosomal HF domain which were shown to form 5 specific heterodimers *in vitro* [51, 57, 151]. With the exception of the Taf3-10 HF domain pair, all other heterodimer partners were found crosslinked (Taf11-13, Taf4-12, Taf6-9, and Taf8-10). TBP crosslinks with several documented interacting partners in the TFIID complex. The N-terminal TAND domain of Taf1 was reported to interact with TBP to inhibit its binding to TATA-box [152]. Taf11 and Taf13 form a dimer and were recently shown to interact with TBP and to compete with the Taf1 TAND domain for TBP binding [29]. The same TBP residues were observed to crosslink with both Taf11-Taf13 and Taf1 N-terminus, indicating that within holo-TFIID, TBP may exist in distinct states and interact with multiple competing partners.

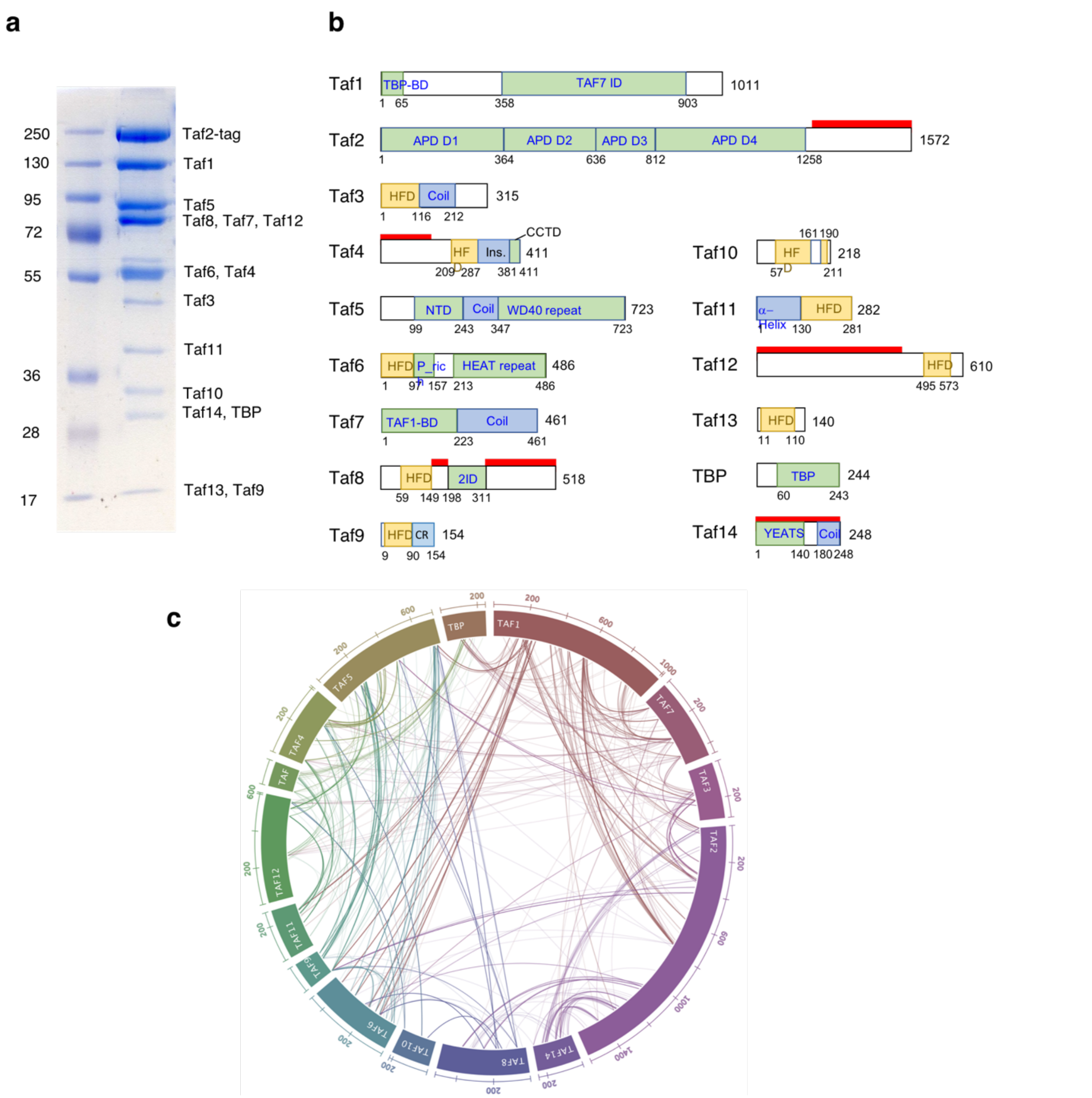


Figure 15: Purification and interaction map of yeast TFIID

a, Colloidal coomassie blue stained SDS-PAGE analysis of TFIID complex purified from the SBP-tagged Taf2 strain. **b,** Schematic representation of the TFIID subunits showing the conserved structural domains (colored boxes) and the yeast specific domains (red bars). Abbreviations: TBP-BD: TBP binding domain (TAND), Taf7 ID: Taf7 interacting domain, APD: aminopeptidase, HFD: Histone fold domain, INS: Insertion, CCTD: conserved C-terminal domain, NTD: N-terminal domain, WD40: structural motif of approximately 40 amino acids, often terminating in a tryptophan-aspartic acid (W-D) dipeptide. HEAT: structural motif composed of two alpha helices linked by a short loop, CR: Conserved region, P-rich: proline rich domain, TAF1-BD: Taf1 binding domain, 2ID: Taf2 interacting domain, CR: conserved HFD flanking region, YEATS: Yaf9, ENL, AF9, Taf14, Sas5 domain. **c,** Subunit-subunit cross-linking map. Line width corresponds to the number of cross-links identified between the two subunits.

Single particle analysis of mildly crosslinked, frozen hydrated *Kp*TFIID molecules resulted in a 3-D reconstruction with an overall resolution of 12.1 Å (**Fig. 16a**). The compact TFIID structure is composed of three lobes interconnected by protein bridges thus forming a closed ring-like system. This arrangement fits with early observations of negative stained particles [3] but differs from cryo-EM studies featuring extended or horseshoe shaped models [25]. Small improvements in various steps of specimen preparation may have stabilized the linkers between the three lobes. To further stabilize TFIID and reach higher resolution, the complex was incubated with TFIIA and the 105 bp long TATA-box containing *K. phaffii* glyceraldehyde-3-phosphate dehydrogenase promoter (pGAP). Electrophoretic mobility shift experiments show that TFIID at a concentration of 0.4 μM interacts with DNA (0.2 μM) and shifts about half of the DNA in the absence of TFIIA. The addition of TFIIA (0.8 μM) not only strengthens the interaction, since the large majority of DNA is shifted, but also modifies slightly the electrophoretic mobility of the complex. Depending of the migration conditions a minor band appears with a faster electrophoretic migration. Mass spectrometry analysis indicated that this minor band contains TBP, Toa1 and Toa2, thus suggesting that the interaction of the TBP/TFIIA/DNA subcomplex with the Tafs is weakened. To probe whether specific promoter elements are involved in the TFIIA-independent recruitment of TFIID to the pGAP promoter, we first mutated the TA-rich elements into G or C in order to degrade the bona fide and any cryptic TATA-box. Electrophoretic mobility shift experiments reveal a slightly lower DNA binding activity in the absence of TFIIA as well as a weaker shift when TFIIA is added as compared to pGAP, but TFIID was still able to bind to the randomized pGAP promoter. In a second experiment, a random region of the coding sequence of the Rpb2 gene was used and TFIID was found to bind similarly to this random sequence as to the mutated pGAP. These experiments show that TFIID binds to non-promoter DNA and that this interaction is stimulated when a TATA-box is present. TFIIA strongly stimulates the binding of TBP to the TATA-box and has a moderate effect in the absence of this promoter element.

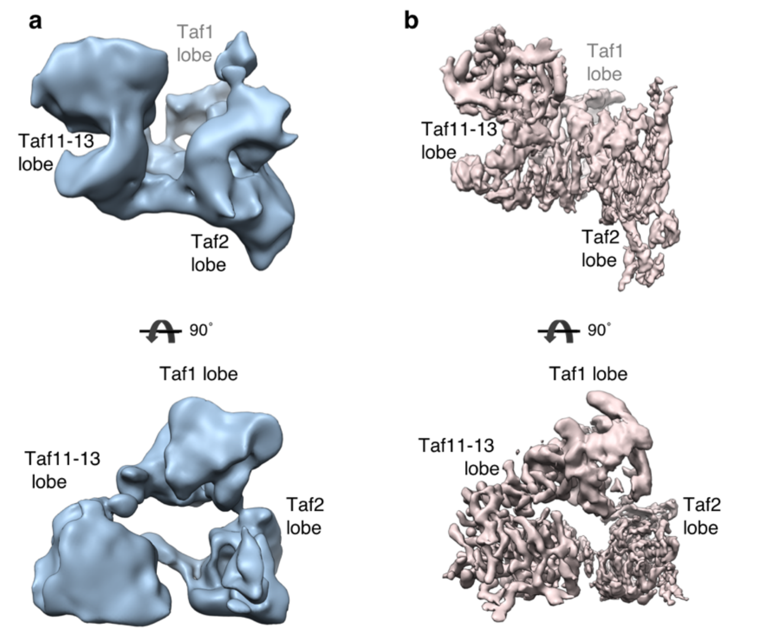


Figure 16: Structural organization of yeast TFIID
a, Cryo-EM model of the yeast *Komagataella phaffii* TFIID at a resolution of 12.1 Å. **b,** Cryo-EM model of the TFIID-TFIIA-pGAP complex.

Single particle cryo-EM analysis of the promoter-bound complex combined with local map refinement showed a significant resolution improvement for the two lobes interacting with DNA while the third one remained poorly resolved (**Fig. 16b**). The overall 3 lobed structure of the TFIID complex was not drastically reorganized upon promoter DNA binding, except for minor conformational changes within the poorly resolved Taf1 lobe. Based on the cryo-EM structures and the crosslinking data, we named the yeast TFIID lobes: Taf1 lobe, Taf11-13 lobe and Taf2 lobe, previously named lobe A, B and C respectively.

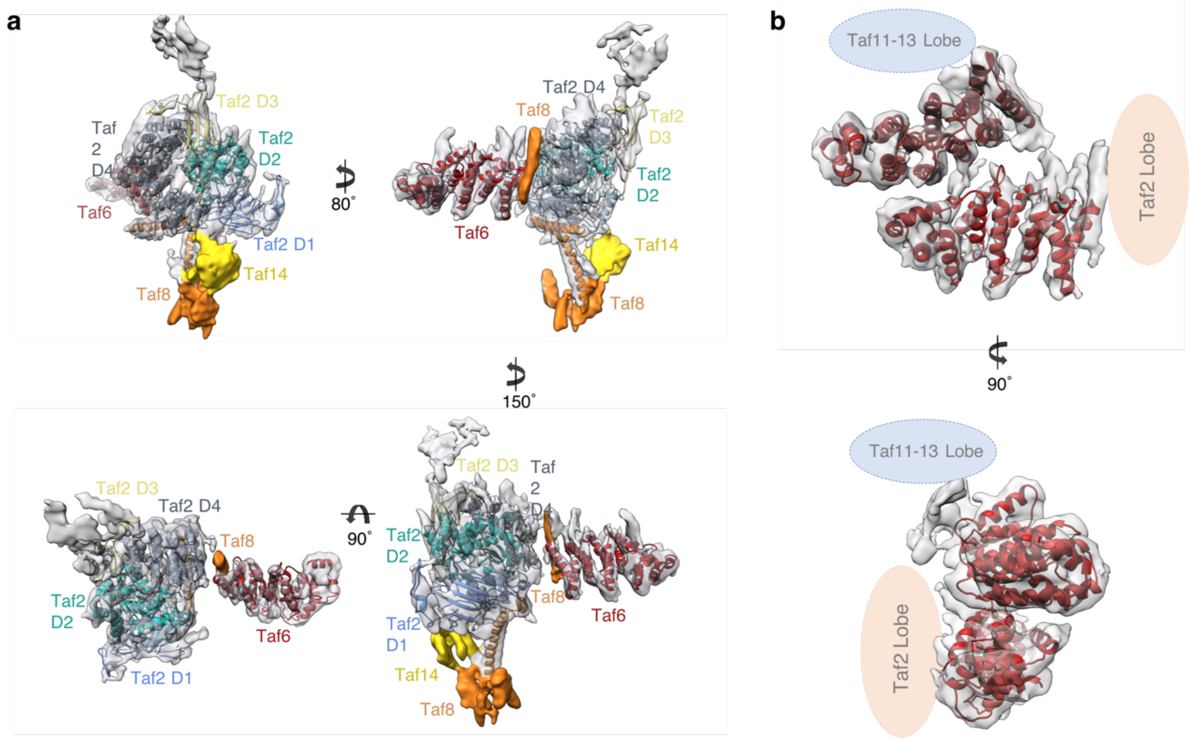


Figure 17: Subunit arrangement within the Taf2 lobe
a, Atomic model docking of the amino-peptidase-like domains of Taf2 (from D1 to D4), the predicted Taf8 helices, and the Taf14 YEATS domain within the Taf2 lobe reconstructed at 4.5 Å resolution. **b,** Linker between the Taf2 lobe and the Taf11-13 lobe consisting of two Taf6 HEAT repeats.

The Taf2 lobe was resolved to 4.5 Å resolution thus revealing the secondary structure of the N-terminal aminopeptidase homology domain of Taf2 (**Fig. 17a**) consistent with the recently published structure of human TFIID [43]. An atomic model of Taf2 residues 36 to 1194, including the 4-subdomain arrangement of the aminopeptidase domain (D1-D4) could be fitted into the cryo-EM map. Recombinant human TAF2 was shown to form a stable complex with the HF domain-containing TAF8 and TAF10 that form a specific heterodimer [51]. The non-HF domain C-terminal part of hTAF8 is crucial to interact with hTAF2. Secondary structure predictions foresee an evolutionary conserved α-helix placed after the HF domain and a short proline rich region (residues 234-259 in *Kp*Taf8). In human TFIID, this helix was proposed to insert between the TAF6 HEAT repeat and TAF2. A helix density is located at the same place in *Kp*TFIID, and this helix crosslinks to both the Taf6 HEAT repeat and to Taf2 D4. Accordingly, its attribution to the conserved Taf8 helix is plausible. The Taf8-10 HF domain dimer could however not be located in the Taf2 lobe density. In most yeast species Taf8 genes have an additional C-terminal extension in which two long α-helices are predicted. Two such long helices are visible in the cryo-EM map at the external surface of Taf2 and their attribution to the C-terminal part of Taf8 is strengthened by their absence in the human TFIID structure [43]. One Taf8 helix contacts two Taf2 helices formed by residues 1071-1081 and 1106-1118, respectively, supporting by multiple crosslinks between Taf2 1068-1104 to Taf8 206-391. The second Taf8 helix contributes to a protein stalk protruding out of the Taf2 lobe and terminated by a protein bulge representing most probably the C-terminal end of Taf8. A compact protein density is observed at the basis of this yeast-specific Taf8 stalk. The size and shape of this density is consistent with the N-terminal YEATS domain of Taf14, a subunit only found in yeast. The YEATS domain is densely crosslinked to the C-terminal end of the aminopeptidase region of Taf2, while the C-terminal coil domain of Taf14 crosslinks to the N-terminal region (1-118) of Taf2. Both regions of Taf2 are located close to the stalk basis. These results agree with a previous study showing that a mutant *Sc*TFIID lacking the yeast-specific Taf2 C-terminal residues 1260-1407 is unable to interact with Taf14 [63]. Furthermore, thermosensitive mutation sites of Taf2, whose phenotype is suppressed by overexpressing Taf14, were also positioned in the vicinity of this density. Altogether these observations indicate that the stalk is formed by Taf14 and the C-termini of Taf8 and Taf2, thus defining a yeast-specific module absent from human TFIID. This module may provide yeast TFIID with a chromatin binding module that was suggested to be replaced in higher eukaryotes by the Taf1 double bromodomain and the Taf3 plant homeobox domain [63].

The linker between the Taf2 lobe and the Taf11-13 lobe is built from two copies of Taf6 HEAT repeats (residues 213-486), thus confirming that Taf6 is present twice in holo-TFIID (**Fig. 17b**). The HEAT repeats are placed at an angle of 40° with a twist of 80° between them and their C-termini are exposed to the periphery of TFIID. Despite their interaction, no symmetry was found between the repeat arrangements. The two-fold symmetry of a recombinant human core TFIID containing two copies of the TAF5, TAF4, TAF12, TAF6 and TAF9 subunits was shown to be compromised by the binding of a single copy of the TAF8-TAF10 heterodimer [26]. The above described putative Taf8 helix, associated with the HEAT repeat closest to the Taf2 lobe, may play a role in introducing asymmetry upon binding to core TFIID.

Connected to the second Taf6 HEAT repeat, the Taf11-13 lobe was resolved to a resolution of 4.8 Å thus allowing the identification of most alpha helices (**Fig. 18a**). A crystal structure of the human Taf5 WD40 propeller and NTD together with the Taf6-Taf9 HF domain heterodimer (courtesy of Dr Imre Berger, unpublished data) fits remarkably well with our cryo-EM structure and helped us to position the WD40 repeat domain of Taf5. Unexpectedly we could identify two additional HF domain heterodimers in the remaining density. The WD40 domain forms the core of the Taf11-13 lobe and coordinates the arrangement of the three HF domain heterodimers (**Fig. 18b**). Two of these were attributed to the Taf6-9 and Taf4-12 pairs since these subunits form a stable subcomplex with Taf5 [26] and are intensively crosslinked to Taf5. Taf11 and Taf13 are the best candidate to form the third Taf5-bound HF domain pair since these subunits strongly crosslink to all partners of this lobe and fail to crosslink with specific subunits of other lobes such as Taf2 or Taf1, except for the highly flexible N-terminal part of Taf1. In the X-ray crystal structure of the human TAF4-TAF12 histone-like heterodimer the α3 helix of the predicted Taf4 HF domain was missing. It was suggested that the α3 helix is separated from the α2 helix by an extended loop and that the characteristic fold would reconstitute with the full Taf4 protein [58]. Our results show that the α3 helix of Taf4 is absent in the TFIID complex and thus confirm the unconventional nature of the Taf4-Taf12 heterodimer. Functional and biochemical analysis of yTaf4 provided strong evidence that the conserved C-terminal domain (CCTD) of Taf4 and the linker adjacent to the histone-fold domain contribute to Taf4-Taf12 heterodimer stability and contains a conserved functional domain essential for yeast growth[151]. The non-attributed densities in the Taf11-13 lobe are likely to arise from Taf12 and Taf4 N-terminal extensions.

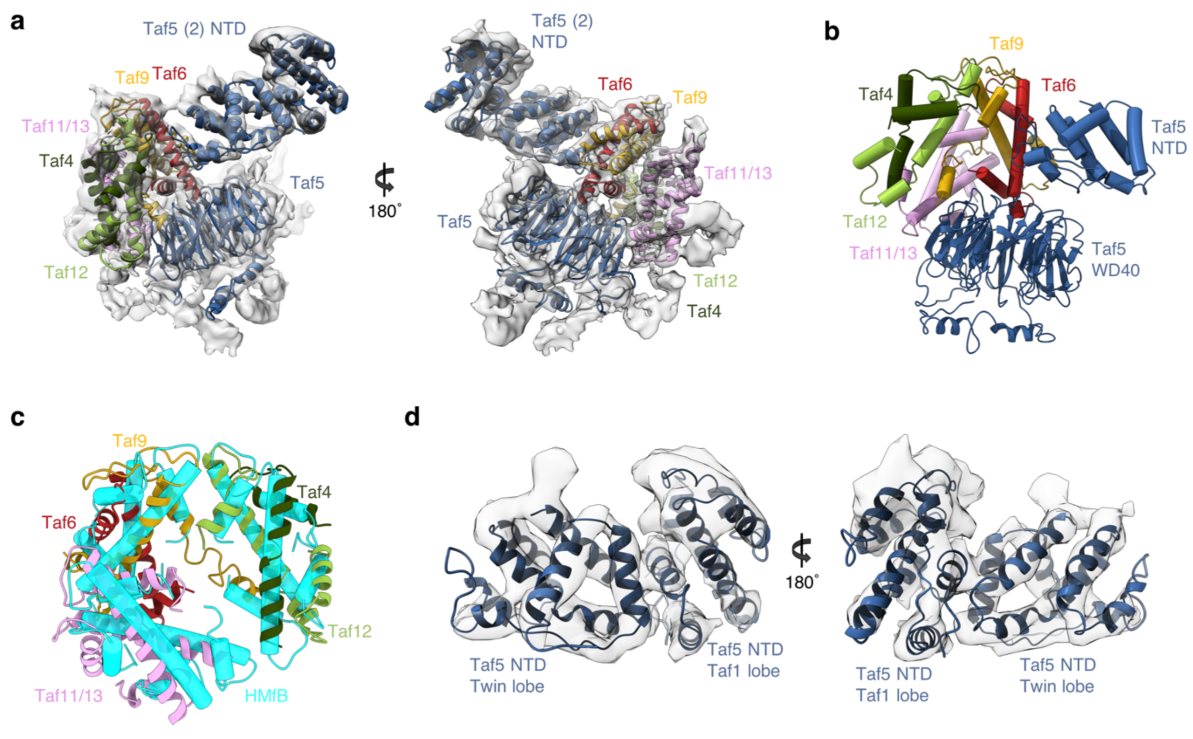


Figure 18: Subunit arrangement of the Taf11-13 lobe
a, Atomic model docking of the Taf5 WD40 repeat, Taf5 NTD, Taf6-9, Taf11-13 and Taf4-12 HF domain heterodimers within the Taf11-13 lobe. **b,** Central role of the Taf5-WD40 repeat in organizing the HF domain heterodimers. **c,** Structural homology between the Taf6-9-4-12-11-13 HF domain hexamer (ribbons) and the archaeal histone hexamer (tubes). **d,** Linker between the Taf11-13 lobe and the Taf1 lobe consisting of two Taf5 NTDs.

The protein density linking the Taf11-13 lobe with the flexible Taf1 lobe is formed by a dimer of Taf5 NTD (residues 99-243) [139] thus confirming that TFIID contains two copies of Taf5[3, 4, 26] (**Fig. 18d**). The two NTD domains are not symmetry related in holo-TFIID but reproduce the arrangement of the asymmetric unit found in crystals of this domain [139].

The Taf1 lobe map shows poor resolution in part due to conformational heterogeneity, which could not be sorted out by local classification or refinement, possibly due to the weak contrast of cryo-EM images. To overcome this limitation, images of the TFIID-TFIIA-pGAP complex were recorded with a Volta Phase Plate (VPP) to produce highly contrasted images [153]. After local alignment, the Taf1 lobe appears as a large globular domain (green and red in **Fig. 19a**) from which an extended protein domain terminated by a bulge protrudes out (grey in **Fig. 19a**). The globular domain contacts the N-terminal Taf2-D1 domain and contains a ring like-structure corresponding most probably to the second Taf5 WD40 repeat as expected by the stoichiometry of Taf5. Since Taf5 and Taf6 are found in two copies in holo-TFIID, we propose that a second Taf5-9-6-4-12 module, similar to the one forming the Taf11-13 lobe, is present in the Taf1 lobe (Green in **Fig. 19a**). We further suggest that a Taf3-Taf10 HF domain heterodimer completes in the Taf1 lobe the hexameric HF domain-structure found around the Taf5 WD40 repeat in the Taf11-13 lobe. Such a position for the Taf3-Taf10 HF pair is strongly supported by the CXML data (red in **Fig. 19a and d**). The VPP image analysis disclosed a large protein density extending from the Taf1 lobe and pointing towards Taf2 (grey in **Fig. 19a**). This flexible arm is also detected without VPP but only when the density threshold is lowered. This domain is likely to correspond to Taf1 and/or Taf7 since the surface of Taf2 facing this domain cross-links preferentially with these two subunits. The crystal structure of a human complex comprising the highly conserved central and amino-terminal fragments of Taf1 and Taf7 [154] was placed in a similar position in the human TFIID complex [43].

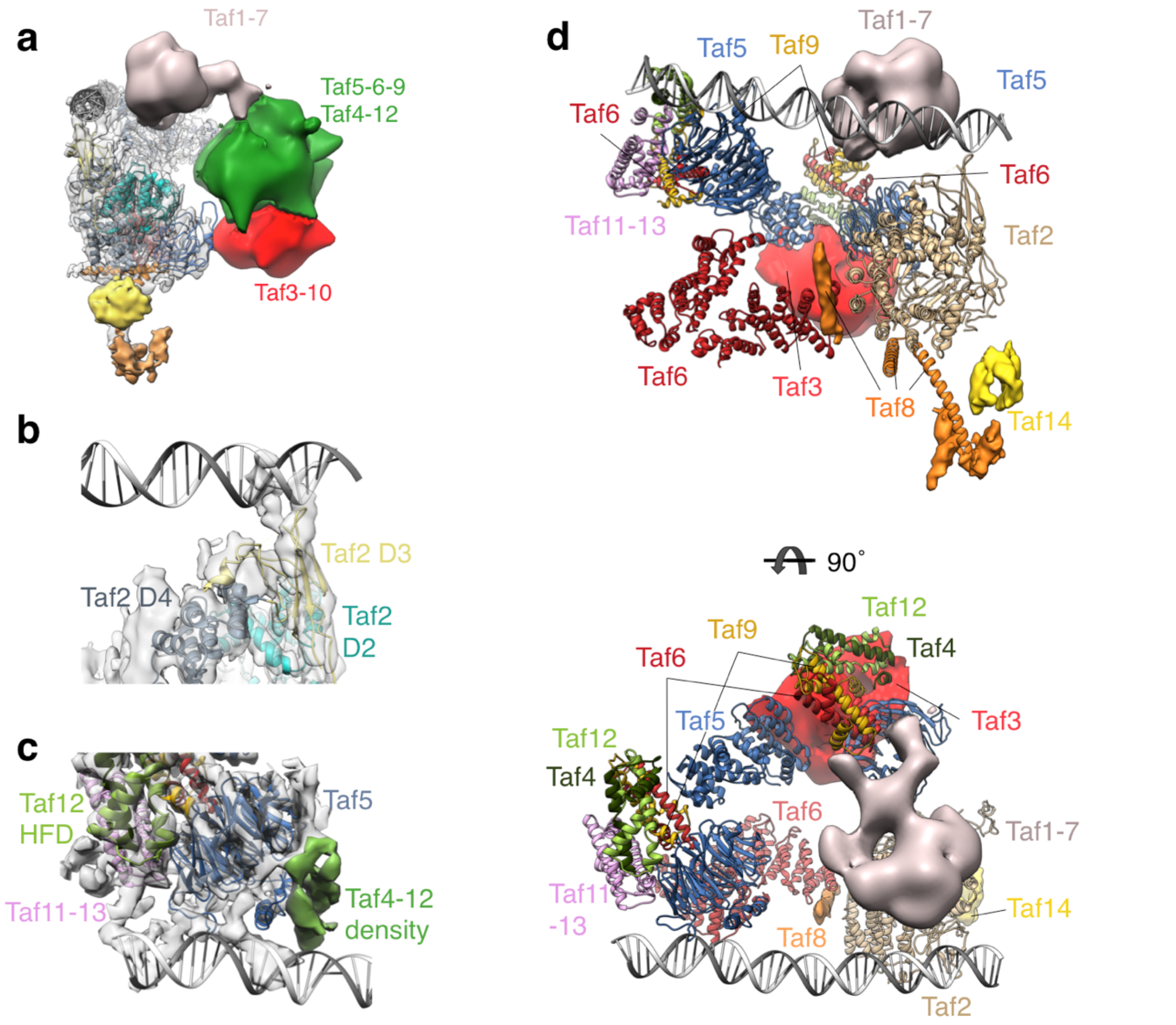


Figure 19: Predicted organization of the Taf1 lobe and DNA interactions
a, Organization of the Taf1 lobe as derived from Volta Phase Plate images of frozen hydrated TFIID-TFIIA-pGAP complexes. The position of the second Taf5-6-9-4-12 module could be determined (green domain). The putative Taf1 flexible domain (grey) is facing the Taf2-bound DNA. **b,** TFIID-DNA interactions within the Taf2 lobe. **c,** TFIID-DNA interactions within the Taf11-13 lobe. **d,** Proposed arrangement of the fitted atomic models and subunits in yeast TFIID.

A continuous thread of density suitable to accommodate a double stranded DNA molecule is located between the Taf2- and the Taf11-13 lobes only in the promoter-bound complex. The signal is weak except at sites were DNA interacts with TFIID, suggesting a low DNA occupancy or an important DNA flexibility. The promoter DNA contacts Taf2 through two arginine and lysine rich loops of the D3 domain of the aminopeptidase domain (residues 733-742 and 646-652) (**Fig. 19b**). The pGAP DNA is clamped between Taf2 and the Taf1/Taf7 arm but the Taf1/Taf7 arm does not interact directly with the pGAP promoter. In the Taf11-13 lobe the DNA path is distant from the HF domain-containing Taf hexamer consistent with the observation that the side chains that mediate contacts between nucleosomal histones and DNA have not been conserved [58]. The DNA interacts with the N-terminal regions of Taf4 and/or of Taf12 consistent with the reported non-sequence-specific *in vitro* DNA binding activities of these subunits [155, 156] (**Fig. 5c**).

Discussion

The present study contributes to our understanding of the molecular architecture of the general transcription factor TFIID by resolving the secondary structure of several Taf subunits. In solution, the yeast complex adopts a compact 3 lobed structure connected by three well resolved linkers. This overall organization is consistent with previous models obtained in negative stain which however showed a gap between the Taf1 and the Taf11-13 lobes [3, 91]. This gap was even more pronounced in our previous cryo-EM map which showed an open, horseshoe shaped arrangement [1, 25] as well as in the human TFIID structure which adopts an extended conformation [2]. This compact organization is maintained by the (Taf6-Taf5)$_2$ tetramers which forms a topologically closed protein ring running through the three lobes (**Fig. 19d**). The more extended conformations that were observed may arise from the disruption of the Taf5-NTD dimerization interaction. This topology provides robustness to the TFIID structure and may explain that the removal of the Taf6-HEAT repeats which connect the Taf2 lobe to the Taf11-13 lobe only moderately affects complex stability [107]. Such a compact architecture of the yeast complex leaves however little space for a major subunit rearrangement between the TFIID lobes, as was described for the human TFIID [64]. Human TFIID was found to adopt two rearranged states in which one lobe (human lobe A) can be associated to either lobe B or lobe C. Furthermore, the presence of both TFIIA and promoter DNA was shown to stabilize one particular rearranged state that enables promoter recognition and binding. In the yeast system, such a rearrangement is not observed and the compact TFIID structure is not affected upon DNA binding.

In vivo self-association studies, quantitative gel electrophoresis profiles and immunoelectron microscopy experiments showed that 7 TFIID subunits are present in more than one copy within the purified TFIID complex. Our structural data confirm directly that Taf5 and Taf6 heterodimer are present in two copies in the native complex. The poor resolution of the flexible Taf1 lobe prevents direct recognition of molecular folds, but we can infer that the HF domain-containing Taf6-9 and Taf4-12 heterodimers are present in two copies. We previously described the dimeric arrangement of a recombinant human core TFIID complex containing 5 human TAFs (TAF5, TAF6-9 and TAf4-12) and presenting a two-fold symmetry [26]. Strikingly, the molecular interactions between core subunits are completely reorganized in the TFIID complex and the two-fold symmetry has been lost. The Taf6 HEAT repeats and the Taf5 N-terminal domains interact in the mature TFIID and form well defined bridges between the lobes, while they are separated in core TFIID. The HF domains of Taf6/9 and Taf4/12 interact with the Taf5 WD40 repeats in full TFIID while this was not the case in core TFIID. The heterotetrameric arrangement of the Taf5-Taf6 core-TFIID subunits adopt an extended circular structure within TFIID that would probably not be stable without interactions with other Tafs, thus supporting the hypothesis that a massive rearrangement takes place upon TFIID maturation. Such rearrangements have been observed upon addition of the TAF8/TAF10 heterodimer to the core TFIID which started to lose internal symmetry [26].

Sequence analysis, *in vitro* interaction data and structural studies showed that nine Tafs contain a histone-like fold allowing the formation of 5 distinct heterodimers and a total of 7 HF pairs when considering the subunit copy number. The analysis of the TFIID density map detected directly 3 heterodimers in the Taf11-13 lobe and predicted 3 pairs in the Taf1 lobe. The predicted Taf10-Taf3 pair could not be confirmed by the current structural analysis due to limited resolution in the Taf1 lobe. Biochemical and structural data suggested a similarity of Taf4 and Taf12 to histones H2A and H2B, and of Taf6-Taf9 to histones H3 and H4, respectively, leading to the proposal that a histone octamer-like structure may exist in TFIID. The cryo-EM structure of yeast TFIID rules out the possibility that Taf4, Taf6, Taf9 and Taf12 form a histone octamer-like arrangement but revealed instead a hexameric arrangement. A similar arrangement was recently described for archaeal histone homodimers [157] where the small basic HMfB proteins, which share a common ancestor with the eukaryotic core histones and are able to interact with DNA by forming a trimeric arrangement of $(HMfB)_2$ homodimers. The structure of three $(HMfB)_2$ dimers and of the HF domain-containing Tafs is highly similar to the nucleosome hexasome, obtained by removing one H2A–H2B heterodimer from the nucleosome structure (**Fig. 18c**). Such an hexameric HF domain-Taf architecture was not reported to assemble *in vitro* thus emphasizing the key role played by the WD40 repeat of Taf5 in holding together the heterodimers.

In metazoan TFIID, Taf1 and Taf2 bind to the conserved initiator (INR) core promoter motif [103, 105], and Taf6 together with Taf9 interact with the downstream promoter element [47]. The binding of specific Tafs to these conserved promoter DNA sequence motifs produces sharp DNase I protections and contributes to a strong and specific interaction of metazoan TFIID to promoters. Neither the INR nor the downstream promoter element have been unambiguously identified in the yeast system [158]. *Sc*TFIID histone fold pairs Taf4-Taf12 and Taf6-Taf9 also display *in vitro* DNA binding activities, but this interaction has not been shown to be sequence-specific [155]. Structural analyses with *Sc*TFIID-TFIIA-activator in complex with promoter-DNA position DNA in contact with the C terminus of Taf2 [25]. However, these interactions of promoter DNA with *Sc*TFIID do not produce sequence specific footprints and the TBP footprint on the TATA-box is predominantly observed [4].

Our results suggest that TFIID has two distinct DNA binding modes. The gel shift experiments show that in the absence of TFIIA, TFIID interacts with DNA whether it contains a TATA-box, a mutated TATA-box or a random coding sequence. Since sequence-specific promoter elements that could interact with Tafs were not identified in yeast, this interaction is likely to be driven by non-specific electrostatic interactions. In these conditions TFIID does not form any specific footprint on DNA and TBP is probably not involved in the interaction since Taf1 negatively regulates its binding to the TATA-box [52, 152]. This TFIIA-independent binding mode is probably reflected by the DNA path observed in the cryo-EM map of the promoter-bound *Kp*TFIID. This path is similar in the human TFIID-TFIIA-SCP complex indicating that the basic DNA interaction modalities are conserved throughout evolution [43]. This binding mode is characterized by two distinct DNA-Taf interaction sites located respectively in the Taf2 lobe and in the Taf11-13 lobe. The Taf6-HEAT linker keeps the Taf2 and the Taf11-13 lobe DNA interaction sites at a constant distance and exposes around 35-40 base pairs (**Fig. 19d**). This bi-partite DNA binding architecture suggests that TFIID could scan the promoter DNA and facilitate the binding of TBP leading to PIC formation or alternatively act as a molecular ruler to select nucleosome free DNA stretches. Yeast TFIID exposes a stretch of 35-40 base pairs of DNA separating the two contact sites. This distance is larger than the yeast 18 bp mean nucleosome linker length [159] suggesting that in this DNA binding mode, the fixed distance could help to select nucleosome free promoter regions and recruit TFIID.

TFIIA modifies the DNA interaction mode of TFIID by competing with the TAND domain for TBP binding thus releasing the inhibitory action of Taf1 [54]. In the presence of TFIIA, a clear footprint has been observed on the TATA-box indicating that TBP can interact with its specific recognition sequence [4]. Our gel shift experiments also show that the addition of TFIIA results in a slower migrating TFIID-DNA species indicating a structural rearrangement of the complex. We could however not detect the promoter-bound TBP-TFIIA sub-complex in our cryo-EM map. We cannot exclude that the binding of TBP to the TATA-box and the resulting bending of DNA may weaken its interactions with Tafs. Such a destabilization was observed in some of our gel shift experiments where a weak, fast migrating DNA band containing TBP and TFIIA was observed. A fragilized TBP-TFIIA-DNA sub-complex may further dissociate upon specimen preparation for cryo-EM. Alternatively, the TBP-TFIIA-TATA-box complex may form while the remaining TBP-less TFIID complex may move along the DNA. In such a situation, the TBP-TFIIA complex would be positioned at variable distances from TFIID and its signal would be averaged out. A similar process was detected in the human TFIIA-TFIID-DNA structure where the promoter-bound TBP-TFIIA complex was readily detected but found located at the periphery of TFIID as if TFIID and the canonical TBP binding site on TAF1 moved downstream by about 30 bp. In metazoan, TAF1 and TAF7 bind to conserved core promoter motive[103, 105], and where shown to form downstream promoter binding module interacting with the Inr motif centered on transcription start site and the MTE and DPE motifs found at +5 to +16 and +17 to +23 respectively [43]. These strong downstream promoter contacts may hold the DNA in place and prevent the Taf-complex from moving further downstream. In yeast, a transcription-dependent recruitment of Tafs to DNA was observed downstream of the TATA-box without TBP or other basal factors [160]. Strong Taf-promoter contacts have not been described and may explain that the Taf complex may move outside of our current EM map.

The DNA-binding Taf11-13 lobe stands out as a key regulatory platform for TFIID function. Taf11 and Taf13 were shown to interact with TBP and compete with the N-terminus of Taf1 for TBP binding [29]. Our CXMS data support such a dynamic association by revealing that the same TBP residues crosslink with both Taf11-Taf13 and the Taf1 N-terminus. Taf4, a component of the Taf11-13 lobe, is crucial for TFIIA binding to TFIID as evidenced by deletion analysis revealing Taf4 sequences next to the HF domain as important for TFIIA-TFIID interaction [161]. A yet to be resolved dynamic interaction network between TFIIA, TBP, subunits of the Taf11-13 lobe and the transcription activators Rap1 is functionally important to regulate the expression of ribosomal genes [162]. In humans, the transactivation domain of the oncogenic transcription factor MYB binds directly to the HFDs of Taf4/12 to drive the expression of genes involved in the development of Acute Myeloid Leukemia [163] suggesting an evolutionary conserved function.

Methods

Preparative scale production of TFIID: The TFIID complex was purified from nuclear extracts of a budding yeast *Komagataella phaffii* strain using a streptavidin-binding peptide (SBP) affinity tag placed at the C-terminus of the Taf2 subunit (Supplementary Fig. 1). 2L of yeast cells were grown at 24°C with glycerol as carbon source and harvested when $OD_{600\ nm}$ reached 12-15. Cells were washed in water and then treated with 10 mM DTT. The cell wall was digested by addition of lyticase and spheroplasts were pelleted at 5,500 g for 20 min. All further steps were performed at 0 to 4°C. Protease inhibitors were added to all buffers. Spheroplasts were disrupted by suspension in a hypotonic buffer (15-18% Ficoll 400, 0.6 mM $MgCl_2$, 20 mM K-phosphate buffer pH 6.6) using a ULTRA-TURRAX disperser. Sucrose (0.1M) and $MgCl_2$ (5 mM) were then added. Nuclei (and some debris) were pelleted at 33,000 g for 37 min, resuspended in a wash buffer (0.6 M Sucrose, 8% PVP, 1 mM $MgCl_2$, 20 mM phosphate buffer pH 6.6) and pelleted again at 34,000 g for 50 min. Nuclei were resuspended in extraction buffer (40 mM HEPES pH 8.0, 300 mM potassium acetate, 20% sucrose, 10 mM $MgCl_2$, 2 mM EDTA, 5 mM DTT) with 20 strokes using a tight pestle in a dounce homogenizer. Following 30 min of incubation, debris were precipitated at 33,000 g for 38 min. The supernatant was collected and 1-2 % PEG 20,000 added in order to precipitate some remaining organelles and membrane parts by a short centrifugation step at 33,000g for 10 min. The PEG 20,000 concentration was then increased to 5.8% and TFIID precipitated in a second short centrifugation step. The pellet was solubilized in a minimal volume and avidin was added to block endogenously biotinylated proteins. The suspension was incubated with streptavidin beads for 4 h in buffer A (40 mM HEPES pH 8.0, 250 mM potassium acetate, 10% sucrose, 2 mM $MgCl_2$, 2 mM DTT) washed 5 times and eluted with buffer A containing 10 mM biotin. The eluate was concentrated with Millipore Amicon-Ultra (50KDa cut-off) and spun in a 10-30% sucrose gradient with buffer B (20 mM HEPES pH 8.0, 150 mM Potassium acetate, 2 mM DTT, 3 mM $MgCl_2$, 0.2 mM EDTA) in rotor SW60 (39,600 rpm for 15.5 h.). TFIID was fractionated at approx. 25% sucrose and concentrated with Amicon-Ultra to ~ 1 mg/ml.

Cross-linking and mass spectrometry: 50 μg of purified TFIID was cross-linked by 3 mM BS3 (Thermo-Scientific) for 2h at 25°C. Samples were digested with trypsin, and the resulting peptides were fractionated by strong cation exchange (SCX) chromatography and analyzed by MS (Orbitrap Fusion). The crosslinked peptides were identified by searching the MS data against a database composed of *K.p.* TFIID subunit sequences using two different algorithms: pLink and in-house designed Nexus (available upon request) as described before [27]. A 5% of false discovery rate (FDR) was used for both pLink and Nexus searches. The circular crosslinking map was generated using ProXL [164].

Formation of promoter-bound complexes: The yeast *S. cerevisiae* TFIIA used for stabilization of TFIID-DNA complexes was recombinantly expressed in *E.coli*, purified from inclusion bodies and reconstituted as described earlier [165].

Promoter DNA fragments were obtained by annealing of equimolar amounts of complementary oligonucleotides at a final concentration of 10μM in 20mM Tris-HCl; 2mM $MgCl_2$, 50mM KCl by heating the mixture to 95°C for 5 minutes and cooling slowly down to room temperature. The 105 nucleotide fragment of pGAP promoter used for EM (5’gacgcatgtcatgagattattggaaaccaccagaatcgaatataaaggcgaacacctttcccaattttggtttctcctgacccaaagactttaaatttaattta-3’) contained 20 nucleotides downstream of the TSS and 40 nucleotides upstream of TATA-box. For the gel-shift experiments fluorescently labeled DNA was used: pGAP (5’-[6FAM]tgtcatgagattattggaaaccaccagaatcgaatataaaggcgaacacctttcccaattttggtttctcctgacccaaagactttaaatttaattta-3’), pGAP promotor in which 17 nucleotides were mutated to delete TATA-box or TATA-like sequences (mut-pGAP: 5’-[6FAM]gactcatgtcatgagatcattggacaccaccagaatcgcgtatcgaaggcgaacacctgtctcacgtctggtgtctcctgacgcacagacttcgaacgta-3’) and a fragment of the coding sequence of the Rpb2 in which two nucleotides were changed to delete polyT or polyA stretches (coRpb2: 5’-[6FAM]gtcttgaccagacaacctgtagaaggtagatcccgtgatggtggtcttcgtctcggagagatggacagagactgtatgattgctcacggt gccgctggt-3').

For the gel-shift experiments 0.4μM TFIID was incubated with 0.2μM of double-strand DNA fragment in presence or absence of twofold molar excess of TFIIA in the buffer containing 15% sucrose, 150mM Potassium acetate, 20mM Hepes pH8.0, 5 mM $MgCl_2$. Protein-DNA complexes were formed for 30 min at room temperature and loaded on a native 1% agarose, 1.5% acrylamide gel containing 5% glycerol and 5mM $MgCl_2$ in Tris-Glycine buffer. Gels were analysed using Typhoon FLA9500 imager (GE Healthcare Life Sciences).

For the EM-studies 0.4μM TFIID was incubated with two-fold excess of TFIIA and 2.5-3-fold excess of pGAP promoter DNA to have all TFIID molecules bound to DNA.

Cryo-EM sample preparation and data acquisition: Freshly purified TFIID or assembled TFIID-TFIIA-DNA complexes were cross-linked with glutaraldehyde (final concentration 0.1%) for 30 min on ice. After the cross-linking reaction was stopped, samples were dialyzed using VSWP MF-membrane Filters (Millipore) to remove sucrose. 3 μl of sample was applied onto a holey carbon grid (Quantifoil R2/2 and UltrAuFoil R1.2/1.3 300 mesh) rendered hydrophilic by a 30 sec glow-discharge in air (2.5 mA current at 1.8x10-1 mbar). The grid was blotted for 2.5 sec (blot force 5) and flash-frozen in liquid ethane using Vitrobot Mark IV (FEI) at 4°C and 95% humidity.

Images were acquired on a Cs-corrected Titan Krios (FEI) microscope operating at 300 kV in nanoprobe mode using the serialEM software for automated data collection [166]. Movie frames were collected in the case of holo-TFIID on a 4k x 4k Falcon 2 direct electron detector at a nominal magnification of 59,000 which yielded a pixel size of 1.1 Å. Seven movie frames were recorded at a dose of 7 electrons per Å2 per frame corresponding to a total dose of 60 e/Å2. In the case of TFIID-TFIIA-DNA the movies were recorded on a 4k x 4k Gatan K2 summit direct electron detector in super-resolution mode at a nominal magnification of 105,000, which yielded a pixel size of 0.55 Å. Forty movie frames were recorded at a dose of 1.32 electrons per Å2 per frame corresponding to a total dose of 52.8 e/Å2, but only the last 38 frames were kept for further processing.

Initial reference generation: Grids containing frozen-hydrated TFIID sample were subjected to tomographic acquisition on a Cs-corrected Titan Krios (FEI) microscope operating at 300 kV in nanoprobe mode using the FEI Tomo software. Images were recorded on a Falcon 2 camera at a nominal magnification of 29,000, which resulted a pixel size of 3.8 Å. Tomographic images were taken with a tilt from -60° to +60° with an increment of 1°. Tomograms were reconstructed in IMOD and sub-tomograms containing single TFIID particles were extracted using the same software [110]. Maximum-likelihood based sub-tomogram alignment and classification was performed in Xmipp.

Image processing: Movie frames were aligned, dose-weighted, binned by 2 and averaged using Motioncor2 [167] to correct for beam-induced specimen motion and to account for radiation damage by applying an exposure-dependent filter. Non-weighted movie sums were used for Contrast Transfer Function (CTF) estimation with Gctf [168] program, while dose-weighted sums were used for all subsequent steps of image processing. After manual screening, images with poor CTF, particle aggregation or ice contamination were discarded. About 6,000 TFIID particles were picked manually using the e2boxer program of EMAN2 [133] and subjected to 2D classification in Relion [169]. Representative class average images showing TFIID in different orientations were then used as references for auto-picking with Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) for both datasets. Several cycles of automatic picking followed by 2D and 3D classification were performed, yielding datasets of 155,620 particles for holo-TFIID. The same procedure was applied to the TFIID-TFIIA-DNA dataset along with random-phase classification (30) resulting in 180,823 particle images. These datasets were analyzed in Relion 1.4 and Relion 2 according to standard protocols. The structures were refined using a low-pass filtered starting model obtained previously by tomography followed by sub-tomogram averaging. Global resolution estimates were determined using the FSC = 0.143 criterion after a gold-standard refinement. Local resolution was estimated with ResMap [170].

Three-dimensional classification of the entire dataset could not clearly separate distinct conformations of TFIID complex. Therefore we carried out a focused refinement of the separate lobes using the masked lobes as references.

Model building: Homology models of protein domains with known atomic structures were made using I-TASSER [171] namely the amino-peptidase domain of Taf2; the HEAT repeats of Taf6, histone-fold domains of Taf4, Taf6, Taf9, Taf11, Taf12 and Taf13, the WD repeats and the N-terminal domain of Taf5. Initial rigid body docking of the homology models into the cryo-EM map of TFIID was performed using ADP-EM [172]. A top scoring solution was found to be in close agreement with previous manual docking. The carbon alpha traces of the models were manually corrected in COOT [173] according to density, taking into account the secondary structure prediction as obtained from Phyre2 [174]. In a few cases (putative part of Taf2 C-terminus, putative long helices in Taf8) alpha helices were placed in a density that was not occupied by the homology models and the helical domain was attributed to a subunit after considering density continuity, 2D predictions, XL/MS and additional published data.

Model geometry was then idealized using phenix.geometry_minimization with secondary structure restraints [175].

All display images were generated using UCSF Chimera [176] and ChimeraX [177].

STRUCTURE AND DYNAMICS OF A 197 BASE-PAIR NUCLEOSOME IN COMPLEX WITH LINKER HISTONE H1

The previously reported crystal structure of the histone H1-bound nucleosome was determined using the isolated H5 globular domain bound to a 167-bp nucleosome bearing the minimal length (10 bp per arm) of linker DNA [86]. The present study focuses on a larger nucleosomal particle, including a full-length linker histone and several helical turns of linker DNA. Structural information on such a particle is needed to localize the non-globular linker histone domains within the nucleosome and to determine the impact of the linker histone on the conformation and dynamics of the linker DNA – information required to better understand the initial stages of chromatin condensation. To this end, we analysed a 197-bp nucleosome containing two 25-bp DNA linker arms and full-length linker histone H1. We determined the structure of this particle by cryo-EM and X-ray crystallography and validated the structure by biochemical analysis. Our results reveal that the H1 globular domain adopts an on-dyad binding mode while the CTD associates primarily with a single linker, strongly disrupting the two-fold symmetry of the nucleosome. These findings advance our understanding of how linker histones associate with nucleosomes and provide an enhanced framework for investigating the assembly of higher-order chromatin structures.

Results

H1 stabilises a compact and rigid nucleosome conformation

We reconstituted nucleosomes using recombinant *Xenopus laevis* or human core histones and a 197-bp DNA duplex comprising the Widom 601 strong positioning sequence [178] or a palindromic derivative (601L) of this sequence [179], respectively. The 601 and 601L nucleosomes were complexed with *X. laevis* histone H1.0b or with a previously described truncation mutant of human H1.5 lacking 50 C-terminal residues [77], hereafter called H1.0 and H1.5ΔC50, respectively. (We use "H1" below to refer generically to both these and other linker histones without specifying the precise isoform or species). Nucleosomes were then analysed by single particle cryo-EM (**Figure 20A-C**). We first determined the structure of 601 nucleosomes lacking H1 and used three-dimensional (3D) classification to sort conformational variants. The cryo-EM reconstruction at 11.4 Å resolution agreed well with the crystal structure of the NCP [66] and allowed fitting of the linker DNA. We observed several distinct linker DNA configurations (**Figure 20A** and **1D**, representative conformations 1-3), which we characterized by measuring the angle formed between each linker and the dyad axis in the planes parallel (angle α) and perpendicular (angle β) to the nucleosomal disc plane (**Figure 20E**). Both angles varied by 25-30°, revealing the highly dynamic character of the linkers, which likely reflects a "breathing" of the histone-DNA interactions near NCP exit. Although the linker arms appear convergent when viewed from the "front" (mean value for α=17.5°± 7.5°), side views show that they diverge from the nucleosomal disc plane (mean β=18.3°± 8.9°), placing the two linker DNA ends far apart (mean separation of 10.1 ± 0.9 nm).

We next analysed H1.0- and H1.5ΔC50-bound nucleosomes. Cryo-EM reconstructions at 11.5 and 6.2 Å resolution, respectively (the resolution difference primarily reflects data collection on different electron microscopes and detectors) revealed a significant change in linker DNA orientation upon linker histone binding, giving particles a more compact appearance (**Figure 20B-D**). 3D classification revealed that the linker orientation was less variable compared to unbound nucleosomes, with narrower distributions for both angles α and β (**Figure 20E**). The α angle shifted towards higher values (mean α= 27.0° ± 3.4° for H1.0 and 25.7° ± 1.7° for H1.5ΔC50), while that for β shifted to lower values (mean β= -0.3° ± 8.7° for H1.0 and 3.8° ± 3.3° for H1.5ΔC50), indicating stabilization of the most convergent linker DNA conformations. Thus, H1.0 and H1.5ΔC50 shift the conformational landscape of the nucleosome to a more compact, rigid state.

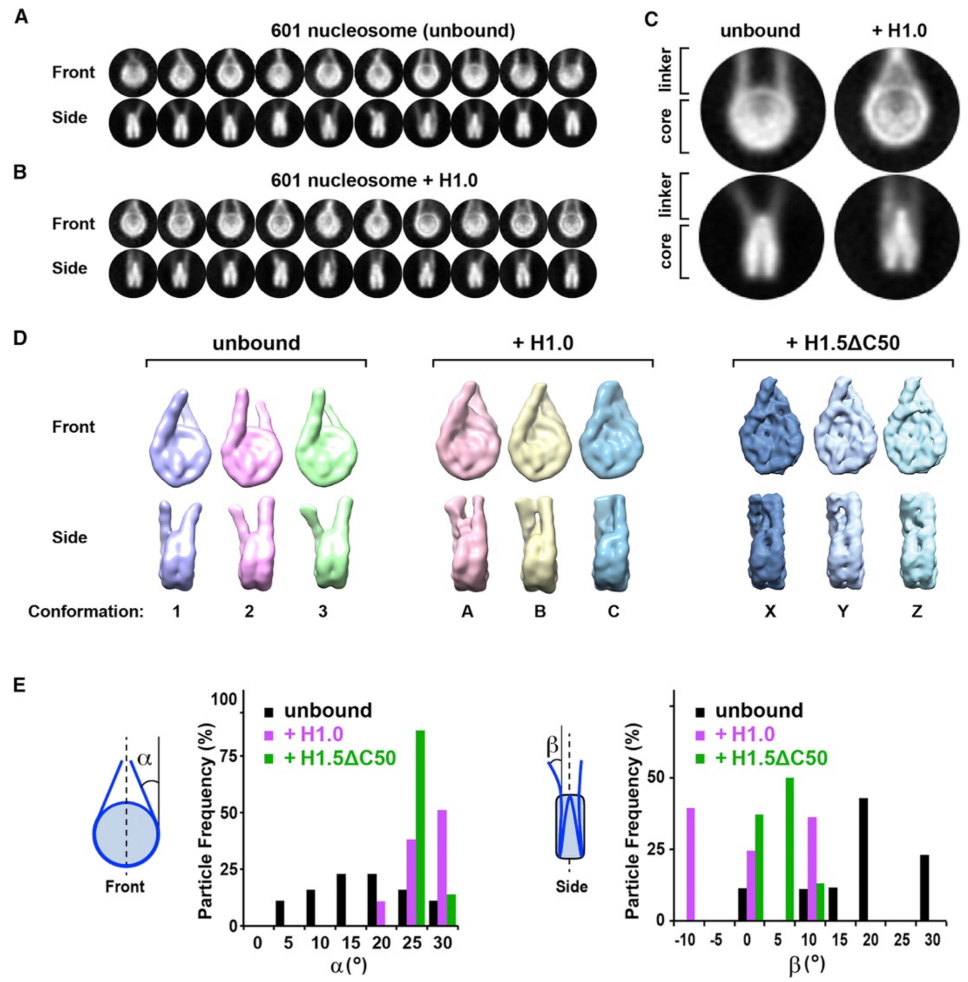


Figure 20. H1 Stabilizes a Compact Nucleosome Conformation
(A and B) Gallery of class averages of 197 bp 601 nucleosomes in the (A) absence and (B) presence of histone H1.0. (C) Close-up views. (D) Representative 3D classes showing different linker DNA orientations in the unbound state (three of eight conformational classes are shown) or bound to H1.0 or H1.5ΔC50 (all three classes are shown). (E) Distribution of linker DNA exit angles in the unbound state (black) or bound to H1.0 (magenta) or H1.5ΔC50 (green).

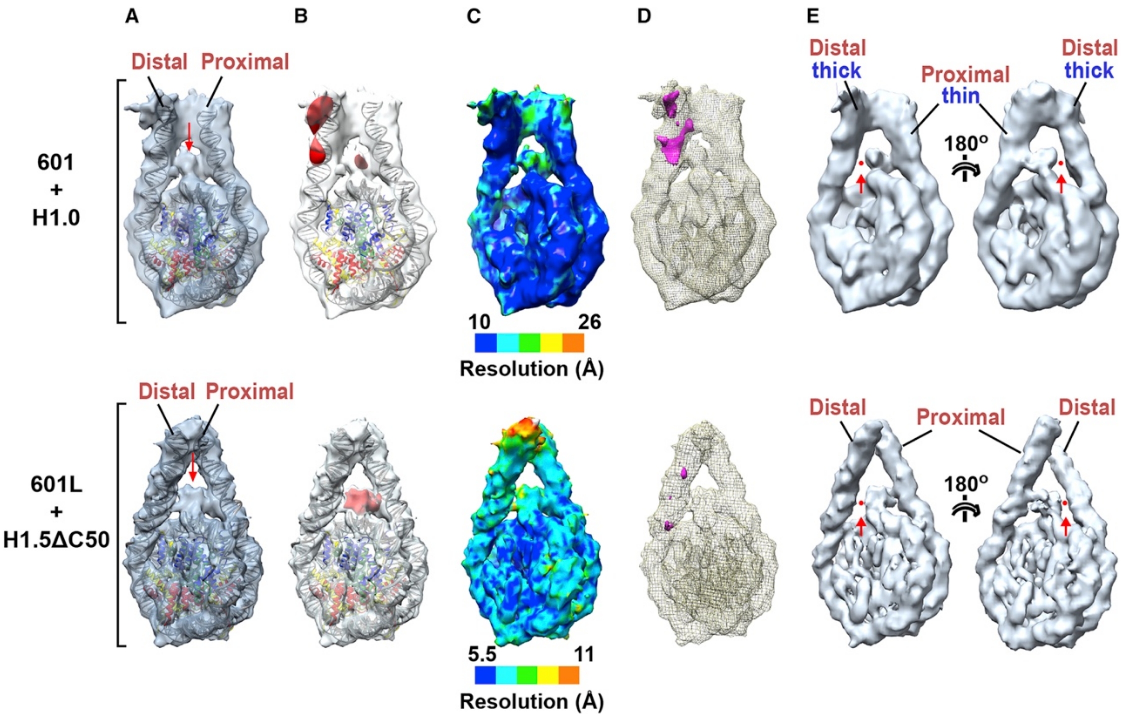


Figure 21. Localization of H1 on the Nucleosome
(A) Atomic models of the NCP and linker DNA fitted into 3D reconstructions of the H1.0-bound 601 nucleosome (top) and H1.5ΔC50-bound 601L nucleosome (bottom). (Structures are of conformations C and Y in Figure 20D.) To highlight H1-occupied density, the H1.0-bound 601 nucleosome map was bandpass filtered to keep spatial frequencies between 10 and 40 Å, while that of the H1.5ΔC50-bound 601L nucleosome was sharpened by applying a negative B factor. The red arrow indicates density attributed to the GH1 domain.(B) Density difference maps (red) calculated between the cryo-EM reconstruction and fitted atomic structures of the NCP and linker DNA.(C) Local resolution maps.(D) Difference map between the two linker arms. The proximal linker density was excised and aligned with the distal linker density. Alignment was performed at a high density threshold to favor the contribution of DNA in linker alignment. A difference map between the aligned linker arms is shown in magenta (threshold, 3 sigma). (E) Views of the H1.0-bound 601 nucleosome (top; bandpass filtered between 10 and 40 Å) and H1.5ΔC50-bound 601L nucleosome (bottom; with B-factor sharpened). Maps are displayed at a higher threshold than in (A)–(D). The red arrow and dot show the loss of contact between the GH1 domain density and one of the linker arms (the thicker distal arm in the case of the 601/H1.0 complex).

Histone H1 confers polarity to the nucleosome

We further analysed the cryo-EM reconstructions obtained for compact H1.0- and H1.5ΔC50-bound nucleosomes (conformations C and Y in **Figure 20D** and shown after high-pass filtering in **Figure 21A**). To localize H1, we compared the cryo-EM density with that calculated from the fitted atomic structures of the NCP and linker DNA. Difference maps revealed additional density on the NCP dyad between the two linker arms attributable to the linker histone globular domains (GH1.0 and GH1.5; collectively referred to as GH1; **Figure 21B**). The local map resolution for these domains (18 Å and 8 Å for GH1.0 and GH1.5, respectively; **Figure 21C**) is somewhat lower than the overall resolution and suggests minor variability in the GH1 domain orientation, possibly reflecting the highly dynamic nature of the H1-

nucleosome association [85]. Fitting the structure of the GH5-bound 167-bp nucleosome [86] into our cryo-EM reconstructions showed a strong overlap between the GH5 domain and the GH1 densities, indicating an on-dyad binding mode for the GH1 domain (confirmed below by crystallography and DNA footprinting). Interestingly, the GH1 domain density is unevenly centered ("lopsided") relative to the nucleosome dyad and appears more intimately associated with one linker than with the other. Indeed, for the more open conformational classes, the GH1-distal linker appears completely detached from the GH1 domain density (**Figure 21E**), revealing at least one of the two GH1-linker interfaces to be unstable.

Strikingly, the difference map calculated for H1.0-bound 601 nucleosomes revealed a second region of additional density on the distal linker (**Figure 21B**, *top panel*), consistent with the thicker appearance of this linker in the original map. The increased thickness is apparently not due to greater linker flexibility, which would have been detected by 3D classification or local resolution measurements (**Figure 21C**, *top*). Indeed, aligning the two linker arm densities and calculating a difference map between them revealed strong positive density indicative of additional mass on the distal linker (**Figure 21D**, *top*). We attribute the extra density to the linker histone CTD. This hypothesis is confirmed by the absence of such density in H1.5ΔC50-bound nucleosomes, consistent with the loss of 50 C-terminal residues in this mutant (**Figure 21B** and **21D***, bottom panels*). This localization of the CTD strongly differentiates the two linker arms, breaking the two-fold symmetry of the nucleosome. Moreover, the many basic residues in the CTD (45 Lys and Arg residues in H1.0) would largely neutralize the negative charge on the CTD-bound linker. Thus, the binding of H1 transforms the nucleosome from a two-fold symmetric particle to one that is strongly polarized both in mass and electrostatic charge distribution.

The H1 globular domain displays an on-dyad mode of nucleosome recognition

We further investigated the structure of the H1-bound nucleosome using X-ray crystallography. Crystals diffracting at 5.5 Å resolution were obtained with a 197-bp palindromic (601L) nucleosome bound to histone H1.0. The structure contains one and a half H1-bound nucleosomes in the asymmetric unit and was solved by molecular replacement using the NCP as a search model. This allowed us to trace the linker DNA and to position the GH1 domain within density. In contrast, density for the N- and C-terminal domains of H1 was poor and therefore not interpreted. The GH1 domain localizes to the dyad axis and interacts with nucleosomal core DNA and with both linkers, exhibiting an on-dyad mode of nucleosome recognition similar to that observed for chicken GH5 bound to the 167-bp nucleosome [86] (**Figure 22A**), consistent with the 80% sequence identity between these two linker histone globular domains. Our crystal structure also agrees well with the cryo-EM reconstruction of the H1.5ΔC50-bound nucleosome (**Figure 22B**), confirming that the H1.0 and H1.5 globular domains adopt the same binding mode. The on-dyad configuration agrees well with previous reports that chromatin association protects several linker histone lysine residues (K52, K55, K69, K82 and K85) from reductive methylation [180], that residue His25 can be cross-linked to terminal regions of nucleosomal DNA [181] and that point mutations of Lys85 enhance susceptibility to micrococcal nuclease digestion [182]. These basic residues all localize close to nucleosomal DNA in our structure.

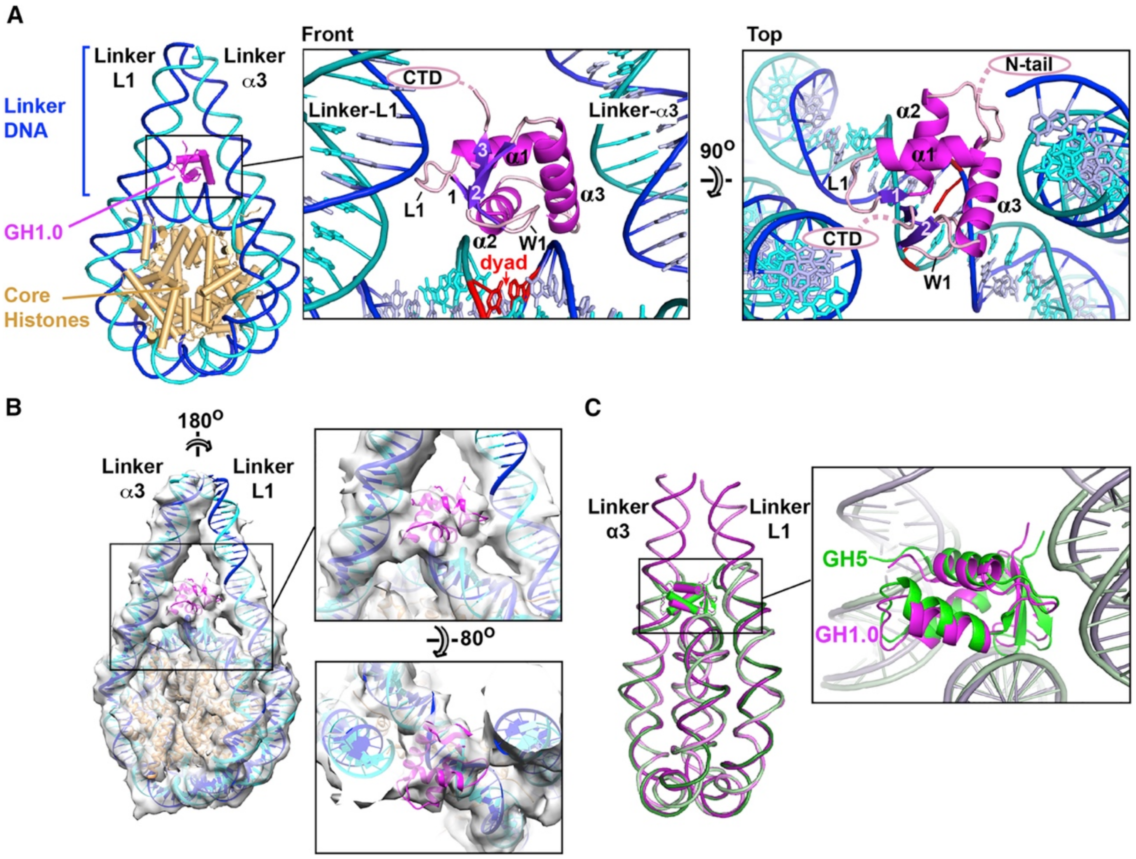


Figure 22. Orientation of the GH1 Domain

(A) Crystal structure showing the GH1.0 domain orientation. The winged-helix fold of GH1 includes a helix-turn-helix (HTH) motif formed by helices α2 and α3 and a "wing" (W1) defined by the β2-β3 loop. The base pair on the dyad axis is in red. (B) H1.0-bound 601L nucleosome crystal structure fitted into the cryo-EM map of the H1.5ΔC50-bound 601L nucleosome. (C) Alignment of the H1.0-bound nucleosome with that of chicken GH5 bound to a 167 bp nucleosome (PDB: 4QLC). The GH1.0 and GH5 domains are related by a 10.5° rotation and by 0.5 Å shift in center of mass.

The linker arms in our H1-bound crystal and cryo-EM structures are farther apart (by ≥5 Å measured half a DNA helical turn from NCP exit) than in the GH5-bound nucleosome structure**,** probably explaining the small (~10°) rotation observed for the GH1.0 domain relative to GH5 (**Figure 22C**). The more open linker conformation results in a contact surface between the GH1.0 domain and the two linkers which is considerably smaller than that observed for GH5 (the buried surface area is 620 Å^{2} for GH1.0 versus 1320 Å^{2}for GH5), rationalizing the disrupted GH1-linker interface observed by cryo-EM (**Figure 21E**). The centre-of-mass of the GH1.0 domain lies midway between the dyad and one of the linkers [linker-α3; linker nomenclature is that of [86]], consistent with the lopsided density observed in the cryo-EM maps (**Figure 21A**). The C-terminal residue of the GH1.0 domain is next to linker-L1 approximately 10 bp from NCP exit (**Figure 22A**), suggesting that the H1 CTD associates with this linker. This attribution is confirmed by fitting our crystal structure into the cryo-EM map for the H1.0-bound 601 nucleosome, which identifies the thicker DNA arm as linker L1. Interestingly, the GH1 domain positions its N-terminal residue next to linker-α3 (**Figure 22A**, right panel), suggesting that the

H1 N-terminal tail may preferentially bind this linker.

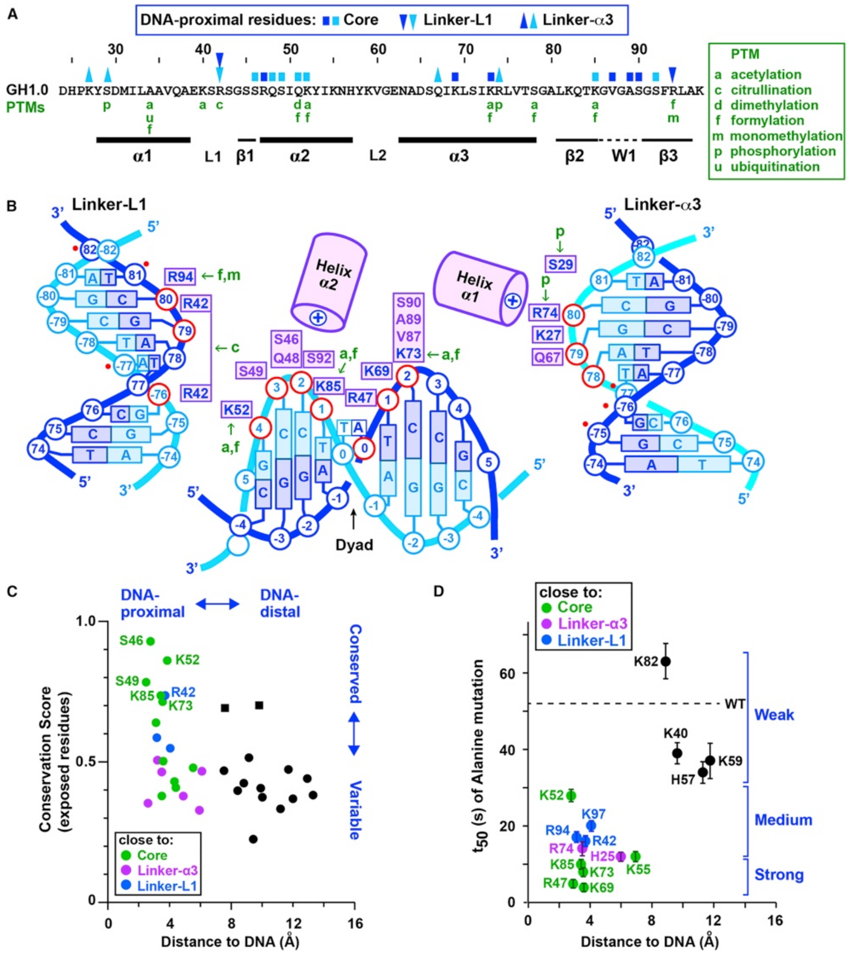


Figure 23. Nucleosome Recognition by GH1.0

(A) Primary structure of the X. LAEVIS GH1.0 domain. Residues close to core or linker DNA are marked by blue (sense) or cyan (anti-sense) squares and triangles, respectively, colored as in (B). Post-translational modifications (PTMs) in mammalian histones H1.1–H1.5 are in green. (B) Summary of DNA-proximal residues. GH1.0 residues are shown next to the DNA phosphate group (in red) to which they are most proximal. Residues shown are within ~4 Å of the DNA, except for Ser29, which is ~5 Å away. Basic residues are in blue, other residues in violet. The six additional linker nucleotide positions contacted by the GH5 domain in the structure are indicated by a red dot. (C) Plot of sequence conservation versus distance from DNA for surface-exposed residues in the GH1.0 domain. For each residue, the distance from each stereochemically allowed rotamer to the closest DNA phosphate atom was measured and the shortest distance was plotted. Residues close to the core DNA or to the α3 and L1 linkers are shown in green, magenta, and blue, respectively. DNA-distal residues are in black. The best-conserved residues localize close to nucleosomal DNA, while most DNA-distal residues are poorly conserved. Exceptions (conserved and DNA-distal; black squares) are Lys40, consistent with an alanine substitution of Lys40 having little effect on stability of the H1-nucleosome complex (see D) and Ser41, which corresponds to an acidic residue in most H1 orthologs. (D) Effect of alanine mutations on half-time of FRAP recovery (T_{50}) plotted versus distance from DNA. FRAP data (mean ± SD) are those of (Brown et al., 2006). Brackets indicate mutations with a strong, medium or weak effect on T_{50}.

The GH1 domain recognizes the nucleosome primarily through the core DNA

The interactions between the GH1.0 domain and nucleosomal DNA are summarized in **Figure 23A** and **23B.** As in the GH5-bound nucleosome structure [86], the GH1.0 domain positions its DNA-proximal residues on the minor groove side of the phosphate backbone, with helix α2 and the W1 "wing" next to the core DNA, helix α3 next to one linker, and loop L1 next to the other. Helices α1 and α2 point their N-termini towards the linker-α3 and core DNA, respectively, stabilizing these two GH1-DNA interfaces via the positive charge generated by the helix dipole (**Figure 22A** and **23B**), reminiscent of the effect observed for helices in the core histones [66]. Nucleotides within contact distance of the GH1.0 domain are disposed nearly symmetrically about the dyad, and include seven nucleotides within the core DNA and three on each linker (**Figure 23B**, red circles). This contrasts with the GH5-nucleosome complex, where the more "closed" linker arm conformation allows GH5 to contact a total of twelve linker nucleotides [86]. Indeed, of the total surface area (1460 Å^2) buried between the GH1.0 domain and the nucleosome, over half (58%) is in the interface with the core DNA, compared to only 24% and 18% for the α3 and L1 linkers, respectively, revealing the core DNA to form the primary binding surface recognized by the GH1.0 domain. This observation is consistent with sequence conservation data: plotting the conservation score derived from an alignment of H1/H5 orthologs onto the GH1.0 domain surface shows that the best-conserved residues localize next to the nucleosomal core DNA, whereas residues next to the linkers are more variable (**Figure 23C**). Also consistent is a study in which the residence time of H1 on chromatin was measured *in vivo* using FRAP [85]. H1 point mutations that reduced residence time localize to DNA-proximal residues in our structure, whereas those that had little effect map to DNA-distal residues (**Figure 22D**). Strikingly, the four mutations with the strongest effect (R47, K69, K73, K85) involve residues located next to the nucleosomal core close to the dyad, whereas mutation of residues adjacent to linker DNA had milder effects. This indicates that, *in vivo,* GH1-linker interactions are weaker than those with the core DNA, consistent with the sizes of the corresponding interfaces in our crystal structure.

H1 adopts an on-dyad binding mode in solution

In order to verify that our crystal structure corresponds to the conformation of nucleosome-bound full-length H1 in solution, we performed site-specific cross-linking and DNA footprinting experiments to confirm specific H1-nucleosome interactions. (Histone H1.0 was used for all experiments unless otherwise specified). Residues Arg42 and Ser66 (located on opposite sides of the GH1 domain next to linkers-L1 and -α3, respectively) were mutated to cysteine (absent from wildtype H1.0) and reacted with 4-azido phenacylbromide (APB), which forms a specific covalent adduct with the cysteine thiol group. Both H1 mutants retain the ability to bind nucleosomes efficiently (e.g., **Figure 24A,** *top*). Radiolabeled nucleosomes incubated with APB-derivatized H1 were exposed to UV radiation, which causes the APB nitrene group to react with nearby nucleotides, generating an H1-DNA cross-link (**Figure 24A,** *bottom* and **24B**). A base elimination reaction and sequencing gel analysis revealed that Cys42 formed cross-links with the half-turn of linker DNA preceding NCP entry (nucleotide positions -80 and -77; **Figure 24C**, orange arrowheads), while Cys66 formed cross-links with the same linker (positions -77 to -74) and with the opposite linker (at approximate positions +75 and +80) (**Figure 24C**, magenta arrowheads). The results are consistent with the GH1 domain adopting two dyad-related orientations corresponding to our crystal structure (**Figure 24D**).

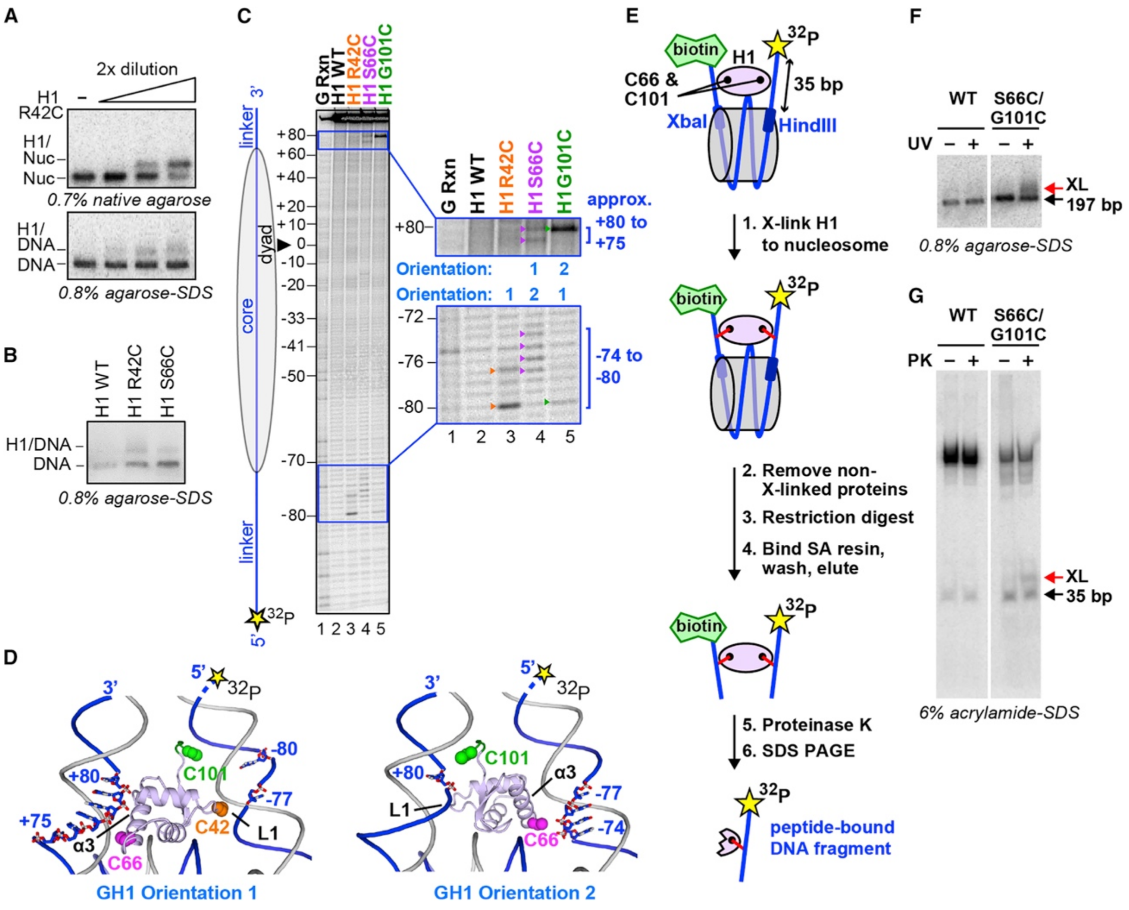


Figure 24. Mapping of H1-Nucleosomal DNA Interactions

(A–D) Site-specific cross-linking of GH1 to nucleosomal DNA. (A) (Top) Native gel showing the binding of APB-derivatized H1 mutant R42C (R42C-APB) to the nucleosome. (Bottom) Denaturing gel showing cross-linking of H1 R42C-APB to nucleosomal DNA after UV irradiation. (B) Denaturing gel showing cross-linking of H1 R42C-APB and H1 S66C-APB to nucleosomal DNA upon UV irradiation. (C) Mapping of cross-linked nucleotides by piperidine base elimination cleavage of the DNA and subsequent sequencing gel analysis. Nucleotides cross-linked to R42C-APB, S66C-APB, and G101C-APB are indicated by orange, magenta, and green arrowheads, respectively. (D) Crystal structure (orientation 1) and dyad-related orientation of GH1 (orientation 2) showing the proximity of GH1 residues to specific linker nucleotides on the radiolabeled strand. Residues 98–101 (green; absent from the crystal structure) were modeled in an extended conformation. (E–G) Simultaneous cross-linking of H1 residues to both DNA linkers. (E) Summary of the cross-linking experiment. Nucleosomes were reconstituted using 5′ biotinylated and 5′ radiolabeled 197 bp DNA containing a specific restriction endonuclease (Xba I and Hind III) site next to each linker arm. (F) APB-derivatized H1 S66C/G101C mutant binds and cross-links in a UV-dependent manner to the 197 bp nucleosome. (G) Elutions with or without Proteinase K (PK) were analyzed on 6% acrylamide-SDS gel, revealing a distinct band (XL) consistent with double cross-link dependent retention of the radiolabeled linker arm.

To verify that the GH1 domain contacts both linker arms in solution, we sought to cross-link this domain to both linkers simultaneously. Because attempts using the H1 double mutant R42C/S66C yielded inefficient double cross-link formation (due to the low efficiency of the individual reactions), we exploited an alternate double mutant, S66C/G101C. Residue Gly101 is located immediately C-terminal to the GH1 domain next to the same linker as Arg42 (**Figure 24D)**. The corresponding Cys mutant yields a highly efficient cross-link (with nucleotide +80, approximately; **Figure 24C**). We reconstituted 197-bp nucleosomes containing a radiolabel on one linker and a biotin tag on the other, each flanked by core DNA bearing a specific restriction endonuclease site (**Figure 24E**). We cross-linked the full-length H1 S66C/G101C mutant to the nucleosome (**Figure 24F**), cleaved the linkers from the core DNA, and affinity purified the biotinylated linker and cross-linked adducts. Proteinase K treatment of the eluted fraction followed by denaturing gel analysis revealed a specific radiolabeled band consistent with H1-mediated tethering of the two linkers (**Figure 24G**). This demonstrates that H1 residues 66 and 101 can simultaneously cross-link to opposite linkers, strongly corroborating our crystal structure.

We next performed hydroxyl-radical footprinting to verify the position of the GH1 domain on the nucleosome core. Both full-length H1 and the isolated GH1 domain make a symmetric footprint on the core DNA, protecting the central base pair plus 3-4 flanking nucleotides on each strand (**Figure 25A** and **25B**; compare lanes 1 and 2 at magenta asterisks). In addition, both H1 and the isolated GH1 domain protect nucleotides within the first turn of linker DNA (**Figure 25A** and **25B**; red and black asterisks) and enhance the protection of core nucleotides in the DNA turn preceding the linkers (**Figure 25A** and **25B**; green asterisks), indicating that H1 induces tighter DNA wrapping around the histone core octamer. These findings recapitulate the footprinting pattern observed for the binding of H1.5 to di- and tri-nucleosomes [77]. We observed the identical protection pattern on dinucleosomes with H1 histones isolated from HeLa cells, as well as with the *X. laevis* oocyte histone B4, an isoform present in early embryonic chromatin which diverges significantly from H1.0 (26% sequence identity overall, 25% in the globular domain). In all cases the observed protection agrees well with the specific protein-DNA interfaces in our crystal structure and with the effect of H1 on linker conformation seen by cryo-EM (**Figure 20D** and **20E**).

While the above footprinting results are consistent with the on-dyad binding mode seen in our crystal structure, they do not formally exclude the possibility that H1 adopts an off-dyad binding mode in solution, since two dyad-related orientations of an asymmetrically positioned GH1 domain can combine to yield a symmetric footprint. To address this issue we prepared nucleosomes lacking either one or the other linker (designated Δlinker-A and Δlinker-B) and confirmed their ability to bind GH1 for linker lengths ranging from 10 to 25 bp (**Figure 25B**). In the off-dyad, single-linker binding scenario, the GH1 domain should bind mono-linker nucleosomes with a preferred orientation and therefore yield distinct patterns of nucleotide protection on the nucleosome core for Δlinker-A, Δlinker-B and the two-linker nucleosome. In fact, incubating either H1 or the isolated GH1 domain with mono-linker nucleosomes yields a footprint on the dyad closely resembling that observed with two-linker nucleosomes (**Figure 25A**, **25B,** magenta asterisks), consistent with the on-dyad binding mode observed in our crystal structure. These data strongly argue against H1 adopting an off-dyad binding mode in solution.

Our crystal and cryo-EM structures show the GH1 domain to be more closely associated with linker-α3 than with linker-L1. Accordingly, H1 should associate with a mono-linker nucleosome by preferentially orienting the linker-α3 binding surface of the GH1 domain towards the single linker. To verify this, we assessed the ability of H1 point mutants S66C and G101C to be covalently cross-linked with mono-linker nucleosomes. Strikingly, whereas both mutants formed cross-links to symmetric two-linker nucleosomes (**Figure 25C**, *bottom*), only S66C was efficiently cross-linked to the mono-linker nucleosome (**Figure 25D**, *bottom*), confirming the greater stability of the GH1/linker-α3 interface. This is consistent with previous observations that mutations on the linker-α3 binding surface of GH5 more significantly reduced nucleosome binding affinity than those on the linker-L1 binding surface [86].

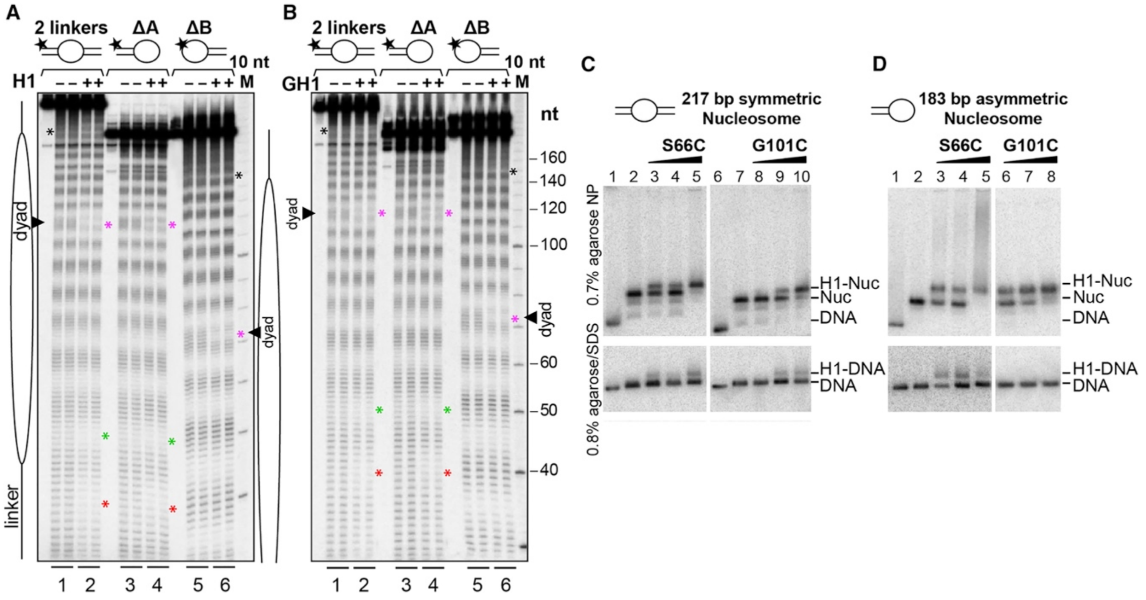


Figure 25. DNA Footprinting and Cross-Linking Analysis of H1 Binding to Symmetric and Asymmetric Nucleosomes

(A and B) Hydroxyl radical footprinting of centrally positioned nucleosomes bearing two linkers (lanes 1 and 2) compared to nucleosomes with only one linker (lanes 3–6). Reactions were performed in the absence (lanes 1, 3, and 5) or presence (lanes 2, 4, and 6) of (A) H1 or (B) the isolated GH1 domain. Cleavage patterns are shown in duplicate. Nucleotide regions protected by H1 or GH1 are indicated by asterisks as described in the text. (C and D) APB-derivatized H1 binding and cross-linking to (C) symmetric, 2-linker nucleosomes, or (D) asymmetric, single-linker nucleosomes. (Top) Native gels showing the binding of H1 S66C-APB and H1 G101C-APB to both (C) symmetric and (D) asymmetric nucleosomes. (Bottom) Denaturing gels showing cross-linking of H1 S66C-APB and G101C-APB to nucleosomal DNA following UV irradiation.

As further validation of our crystal structure, we performed a molecular docking analysis to identify the most probable GH1 domain orientation in solution compatible with the above biochemical data. We docked the GH1 domain to the 197-bp nucleosome by two approaches. In a data-driven approach using the program HADDOCK [183], we used the above cross-linking and footprinting results as interaction restraints to guide the docking procedure. In certain docking experiments we also included previously reported data indicating close proximity of specific residues (His25 and Lys85) to DNA [180-182] as additional restraints. Using either the full set of restraints or various partial subsets, the best-scoring solutions consistently displayed an on-dyad binding orientation which clustered around the GH1 domain orientation observed in our crystal structure. In a separate, unbiased docking approach, we used the program Autodock Vina [184] to generate energetically favoured GH1-nucleosome configurations which were subsequently screened for consistency with the biochemical data. This approach identified a single solution similar to the configuration observed in our crystal structure. Thus, both molecular docking approaches support an on-dyad binding mode for H1 in solution.

Discussion

In this study, we used structural and biochemical techniques to investigate an intact 197 bp nucleosome containing full-length histone H1. Our cryo-EM analysis shows that H1 binding induces the nucleosome to adopt a more compact conformation with reduced linker arm flexibility. This is significant because a more homogeneous nucleosome conformation would likely facilitate assembly into a regular helical structure and promote condensed fibre formation. The binding of full-length and C-terminally truncated H1 constructs yielded similar effects on linker conformation and dynamics, suggesting that much of the CTD is dispensible for inducing a more compact and rigid nucleosome structure. The linker arms in our H1-bound crystal and cryo-EM structures are farther apart than in that of the chicken GH5-bound nucleosome [86], resulting in relatively small GH1-linker interfaces. Our cryo-EM data show that dynamic flexibility of the linker DNA can lead to increased linker separation and to the disruption of one of the GH1-linker interfaces, yielding a “two-contact” binding mode in which only a single linker and the core DNA interact with the GH1 domain (**Figure 21E**). This interdependence between linker separation and the size of the H1-nucleosome interaction surface suggests how factors affecting the exit/entry angle of linker DNA could modulate the stability of H1 binding. For example, the binding of linker histones is abrogated by the defective docking domain in the H2A histone variant H2A.Bbd, which causes the unwrapping of ~10-15 bp at each end of the NCP [185]. Dynamic linker flexibility may also contribute to the ability of transcription factors to compete with H1 to bind cognate sites located within the linker DNA [186].

A striking result of our study is the observation that the highly basic CTD of H1 associates primarily with a single linker, whereas the GH1 domain is positioned more closely to the opposite linker. This arrangement confers a notable asymmetry to the nucleosome, both in mass distribution and electrostatic character, as the highly basic CTD would neutralize the negative charge of the associated DNA linker. The lopsided positioning of the GH1 domain relative to the dyad and to the two linkers also contributes to the particle’s asymmetry. Such asymmetry in H1-bound nucleosomes is likely to have significant consequences for the formation of higher-order chromatin structures. For example, in a nucleosomal array with a two-start helical configuration, H1 proteins bound to adjacent nucleosomes with the same (head-to-tail) polarity would yield a different spatial arrangement of CTD-bound linkers than would proteins bound with opposite (head-to-head) polarity, resulting in distinct mass and electrostatic charge distributions (**Figure 26A**). These two configurations are characterized by different repeating structural units (dinucleosome versus tetranucleosome) and could conceivably stabilize different higher-order chromatin conformations. On a more speculative note, chromatin assembled with different H1 subtypes has been reported to exhibit distinct nucleosomal spacing [187]. The CTD is responsible for much of the heterogeneity between H1 subtypes and may therefore be an important determinant of nucleosomal repeat length. The observation that the CTD associates primarily with one linker suggests how nucleosomal spacing might be influenced by this domain: the affinity of core histone proteins for the CTD-bound stretch of linker DNA would be reduced relative to distal (more negatively charged) naked DNA, thereby favoring a minimal length of DNA between neighbouring NCPs.

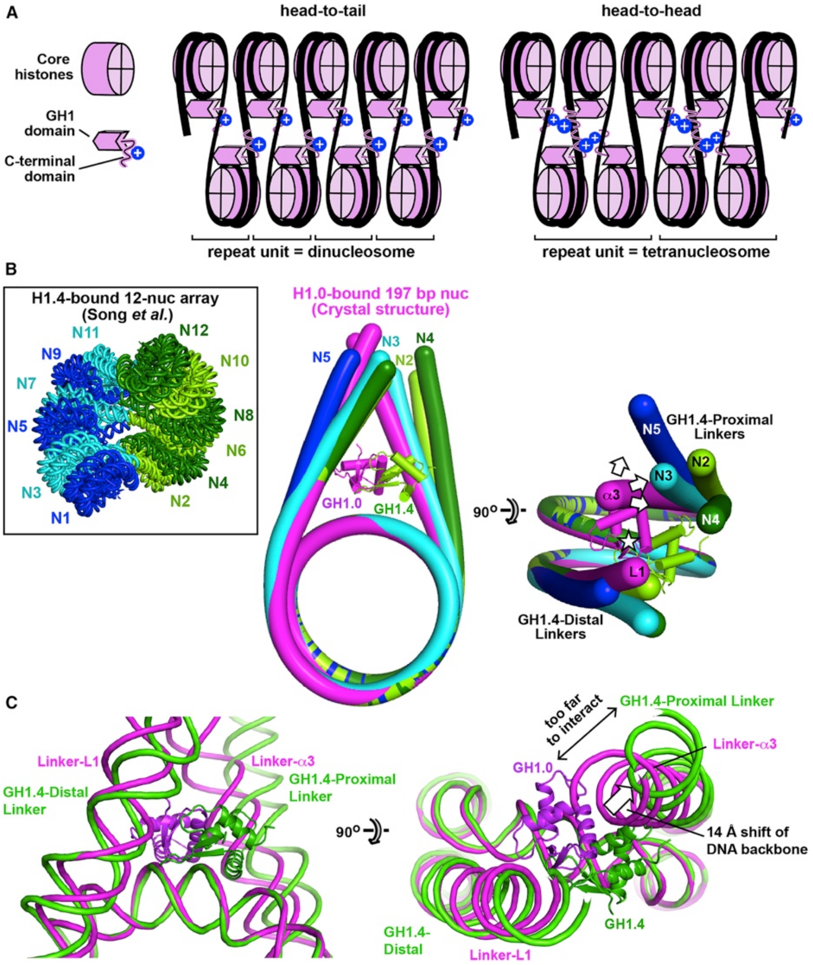


Figure 26. Implications for Higher-Order Chromatin Structures
(A) The asymmetric localization of the CTD may influence the assembly and properties of higher-order structures. Two hypothetical arrangements shown for H1-bound nucleosomes within a two-start helical array give rise to distinct mass and electrostatic charge distributions. (B) Comparison of linker arm geometry with that observed in the condensed 12-nucleosome array of Song et al. (2014). Nucleosomes N2–N5 of the 12-nucleosome array were aligned onto the H1.0-bound 601L nucleosome crystal structure (complex A) by superimposing the nucleosomal cores. The DNA from the crystal structure is in magenta, while that for N2–N5 is in lime, cyan, dark green, and blue, respectively. (Only four nucleosomes of the array are shown, because the three tetranucleosomal units have similar conformations. N5 is shown instead of N1, because the latter lacks the first linker arm.) The GH1.0 domain from the crystal structure and the GH1.4 domain bound to N2 are also shown. The asterisk indicates the pseudodyad axis. The arrows show the displacement of GH1.4-proximal linkers relative to linker-α3 of our crystal structure. The mean displacement of the DNA backbone measured one helical turn from NCP exit is 14.5 ± 6.3 Å between the GH1.4-proximal linkers and Linker-α3, and 4.0 ± 1.6 Å between the GH1.4-distal linkers and Linker-L1. (C) Comparison of linker arm geometry between the H1.0-bound crystal structure and nucleosome N2 of the 12-nucleosome array

Our structural data reveal the GH1 domain to be in contact with both linkers and with the nucleosome dyad, similar to the on-dyad configuration reported for chicken GH5 [86]. Cryo-EM, crosslinking, and footprinting analyses confirm that this binding mode also occurs in solution. We observe the same on-dyad binding mode for the globular domains of both *Xenopus* H1.0 and human H1.5. These two domains share 47% sequence identity and differ at numerous (23 out of 35) solvent-exposed residues, indicating that even considerably divergent H1 isoforms can adopt the same binding configuration. By contrast, the *Drosophila* GH1 domain (43-46% identical to chicken GH5, *Xenopus* GH1.0 and human GH1.5) has been observed to bind off the dyad [87]. Moreover, a chicken GH5 mutant could be engineered to adopt an off-dyad binding mode by replacing five surface-exposed residues with the corresponding *Drosophila* GH1 residue, confirming that GH1 sequence variation can modulate binding configuration [88]. Interestingly, human GH1.5 matches *Drosophila* GH1 at two of these mutated positions, raising the possibility that the H1.5 on-dyad configuration may be less stable than that of the H1.0 isoform.

On a related note, the on-dyad binding configuration observed for the globular domains of chicken H5 [86], *Xenopus* H1.0 and human H1.5 (this work) bound to a mononucleosome differs markedly from the off-dyad binding reported for the human H1.4 globular domain in condensed 12-nucleosome arrays [89]. This is striking because H1.4 and H1.5 are closely related paralogs (95% sequence identity within the globular domain) and the few divergent residues are unlikely to account for the different binding configurations. How then do the different binding modes arise? Aligning the individual nucleosomes of the 12-nucleosome array with our H1-bound crystal or cryo-EM structures reveals substantial differences in linker arm conformation: whereas the linkers in our structures are essentially symmetrical relative to the nucleosome dyad, those of the condensed array show a much greater degree of asymmetry (**Figure 26A**). This is due to the twisted fibre geometry of the array, which requires the two linkers of each nucleosome to follow non-superimposable trajectories as they connect to the preceding and subsequent nucleosome. Consequently, whereas the GH1.4-distal linkers of the array superimpose reasonably well with either of the two linkers in our crystal structure, the GH1.4-proximal linkers do not (see **Figure 26A**, legend). The latter linkers are displaced away from the pseudodyad axis, too far to interact with a GH1 domain bound on the dyad (**Figure 26B**). Thus, the DNA conformation in the condensed array would significantly destabilize the on-dyad configuration, as a GH1 domain adopting such a binding mode could at best interact with only one linker, not two. Indeed, the observed GH1.4 domain in the array adopts a completely different orientation (rotated by 85°) and is substantially shifted (by ~20 Å) relative to the on-dyad GH1 orientation, presumably so as to optimize interactions with a single linker and the core DNA. Because our crystal and cryo-EM mononucleosome structures likely represent nucleosomes in the uncondensed state, the above findings suggest that (at least for histones H1.4 and H1.5) chromatin condensation is associated with a switch from on-dyad to off-dyad binding. More generally, these findings suggest that a causal relationship may exist between linker conformation and GH1 binding mode, and consequently that different GH1 binding configurations might be associated with distinct higher-order chromatin structures.

A number of post-translational modifications (PTMs) have been reported for the globular domain of mammalian somatic linker histones [188-190]. Interestingly, most of these occur on DNA-proximal residues and are predicted to destabilize the H1-nucleosome complex (**Figure 23A** and **23B**). For example, phosphorylation has been observed on Ser29 (H1.0 numbering) in histones H1.1-H1.4 from multiple murine tissues, and on a serine corresponding to H1.0 residue Arg74 in mouse kidney histones H1.2-H1.4 [188]. Phosphorylation at these sites would cause a strong electrostatic repulsion with the linker-α3 phosphate backbone. Citrullination of Arg42 (H1.0 numbering) in histones H1.2-H1.5 in mouse pluripotent stem cells has been linked to chromatin decondensation and to the enhanced expression of genes involved in stem cell development and maintenance [190]. The loss of positive charge induced by citrullination would weaken the interaction of Arg42 with linker-L1 and promote H1 dissociation. The interaction between GH1 and linker-L1 would similarly be destabilized by the formylation of Lys106 in histone H1.2 (corresponding to Arg94 in H1.0) observed in murine seminal vesicles [188]. Likewise, the acetylation or formylation of three lysines (K52, K73 and K85 in H1.0) located next to nucleosomal core DNA in histones H1.1-H1.4 in human cell lines and in H1.2-H1.5 in various murine tissues [188, 189] would favour the eviction of H1 from the nucleosome. Our structural data thus provide molecular insights into how the post-translational regulation of histone H1 is achieved.

In conclusion, our structural and biochemical results paint a coherent picture of how histone H1 interacts with a ~200-bp nucleosome. These results advance our understanding of nucleosome recognition by linker histones and will inform future efforts to elucidate the mechanism of chromatin condensation and the architecture of higher-order chromatin structures.

FUTURE PROJECT

The main focus of my research was on the general transcription factor TFIID after my graduation. TFIID is part of the transcription preinitiation complex where it ensures the proper place of the transcription. Despite the lot of effort up to date, the function and exact structure of TFIID is not known. There are several intriguing open questions, for example what is the role of the Tafs in DNA recognition; whether TFIID just loads TBP onto DNA and leaves, or it has other roles as well during transcription initiation; how does it interact with the other factors in the transcription preinitiation complex. The structure of the transcription preinitiation complex was described *in vitro*, but only with the TBP subunit of TFIID. How the Tafs are involved in the complex is not known. A lot of evidences point towards the existence of promoter-specific assemblies *in vivo*.

To elucidate the role of TFIID in transcription preinitiation complex formation, I started collaboration with Dr. Laszlo Tora and Dr. Jeff Ranish, which I will describe in the following pages.

Eukaryotic gene expression requires the assembly of the transcription preinitiation complex (PIC) on active gene promoters. The role of this ~60 protein complex is to position accurately the RNA polymerase II on transcription start sites. In the last decades, the PIC components have been identified and structures of the *in vitro* reconstituted core PIC has been solved. Despite intensive efforts, information on endogenous PIC composition and its structural variability is scarce. In particular no information is available on PICs containing the general transcription factor TFIID and on possible promoter specific assemblies.

Our project aims to determine a high-resolution structure of a reconstituted TFIID-containing PIC by single particle cryo-EM. Preliminary results on TFIID bound to a TATA-box containing promoter revealed the structure of TFIID at an unprecedented resolution of 4.7 Å, except for a very flexible lobe. A key element to reach such a high resolution was the development of a novel purification protocol that preserves the integrity of TFIID. We plan to determine the structure of reconstituted functional TFIID complexes on natural yeast promoters containing or lacking the TATA-box in complex with recombinant TFIIA or TFIIB to better understand its role in PIC formation.

The second objective of this proposal is to study the subunit composition and the molecular structure of promoter bound TFIID-containing initiation complexes assembled *in vivo*. When our novel purification protocol is used in conditions where nuclear DNases are activated we were able to purify DNA-bound TFIID complexes and preliminary proteomic analysis showed that these complexes also contained Pol II, general transcription factors, coactivators and chromatin remodellers. This unique finding opens the possibility to determine the structure of native transcription initiation complexes, their interaction landscape, to analyse their DNA content and to correlate this information with genome-wide GTF occupancy and gene expression profiles. This analysis will identify stable transcription factor assemblies that are involved in transcription initiation and our preliminary data indicate that they differ significantly from *in vitro* reconstituted PIC assembly. Our results will shed new light on the transcription initiation process by determining the composition and structure of key TFIID-containing initiation complexes formed in live cells.

STRUCTURE AND FUNCTION OF ENDOGENOUS TFIID-CONTAINING TRANSCRIPTION PREINITIATION COMPLEX

Our project is highly innovative for three major aspects:

First, the exact structure of holo TFIID is not known, and its precise role in PIC formation is only poorly understood. Due to the difficulty of purifying endogenous TFIID in large amounts, and due to its intrinsic flexibility, the existing TFIID structures are either at low resolution, or only the arrangement of a small subset of subunits has been described at near-atomic resolution [25, 43]. To overcome these limitations and to investigate TFIID-containing endogenous PICs on natural promoter DNA, our laboratory is developing purification methods to obtain intact endogenous TFIID complexes from the yeast *Pichia pastoris* in sufficient amount to perform electron microscopy studies by tagging several of its subunits.

Second, the structure and composition of endogenous TFIID-containing complexes assembled *in vivo* on natural yeast promoters have never been studied so far. The only structural information available comes from a PIC assembled *in vitro* on an artificial TATA-box containing promoters and lacking TFIID. In our preliminary experiment, proteomic analysis of purified native TFIID complexes showed that subunits of major transcription complexes such as Pol II, Mediator, GTFs and chromatin remodelers co-purified with TFIID. Some of these

TFIID-containing supramolecular assemblies could be purified by a second affinity tag placed on TFIID's partner and were shown to contain a DNA component.

Third, by analyzing the DNA co-purified with endogenous TFIID-containing complexes and by using tandem-affinity purification with tags on TFIID and on different other members of the PIC we will have the possibility to define the promoter selectivity of different TFIID-containing complexes and correlate them with genome-wide active promoter categories.

Preliminary data: To obtain high amounts of transcription complexes we started to use *Pichia pastoris* (Pp), as this yeast can be cultured to very high cell density by maintaining an exponential growth rate. Our preliminary results show that we could get large amounts of TFIID (Figure 1) by tagging the Taf2 subunit of PpTFIID. This anti-Taf2 affinity purification allowed us to obtain a highly homogenous TFIID= sample, which we can use for structure determination by cryo-EM (Figure 1). Probably in part due to its intrinsic flexibility, the structure of TFIID could only be solved at a resolution of 12.1 Å, which is not sufficient to detect secondary structure elements. We explored possibilities to reduce this flexibility by including interaction partners to stabilize the complex. We obtained a preliminary structure of promoter-bound TFIID-TFIIA ternary complex at sufficiently high resolution to identify α-helices and secondary structure elements. We could identify around two-third of the subunits in the cryo-EM structure (Figure 2). We started the investigation of the subunit interactions of TFIID alone with J. Ranish using crosslinking coupled with mass-spectrometry (XL-MS) to provide a linkage map of the complex and to study the possible rearrangement of the complex upon DNA binding (Figure 3).

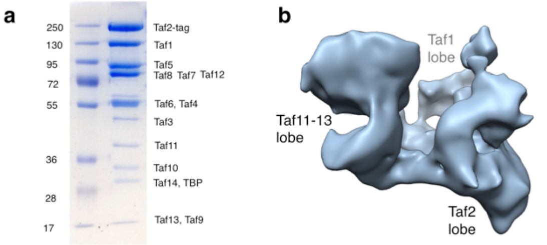


Figure 1 *Pichia pastoris* TFIID purified by SBP tag on Taf2. **(a)** Colloidal coomassie blue stained SDS-PAGE analysis of the subunit composition of TFIID complex. (b) Preliminary cryo-EM model of TFIID at a resolution of 12.1 Å.

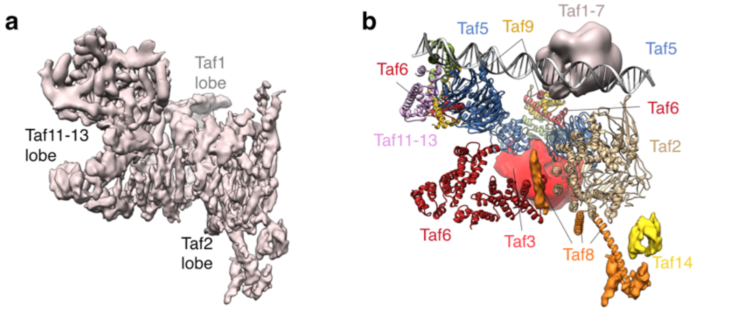


Figure 2 Preliminary Structure of the promoter-bound TFIID
(**a**) Preliminary cryo-EM model of the TFIID-TFIIA-pGAP complex. (**b**) Based on our preliminary results, we fitted atomic models and subunits in yeast TFIID.

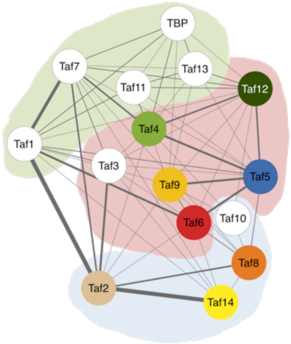


Figure 3 Subunit-subunit cross-linking map of TFIID.

In addition, we implemented additional modifications to the purification protocol of endogenous TFIID, to preserve its interactions with its cellular partners forming the PIC. To purify promoter associated TFIID-containing transcription complexes we tagged a subunit of Pol II in addition to Taf2 in TFIID. By carrying out a double tag purification protocol, we obtained very promising preliminary results showing by mass-spectrometry that all known components of the PIC are present in the isolated sample. We also found that these complexes contain DNA, probably promoter fragments, opening the door to study their promoter occupancy. In addition, we obtained amounts suitable for structural studies using cryo-EM (Figure 4). To obtain DNA-bound PICs we induced the activity of endogenous DNases in the cell during purification. This opens a completely novel opportunity to study **intact endogenous preinitiation complexes**. With the participation of the labs of László Tora, an expert in eukaryotic transcription, and that of Jeff Ranish, a specialist in cross-linking mass-spectrometry analyses and our lab, a specialist in cryo-EM, a multidisciplinary consortium will be formed, which gives the unique possibility to study the endogenous TFIID-containing PICs from the yeast *Pichia pastoris* to elucidate endogenous transcription initiation processes.

Methodology and risk management

For sake of clarity, detailed risk assessment and elaboration of alternative options are described at the level of each task. We have produced key preliminary results, which show that several aspects of the project are feasible and that we will undoubtedly provide the scientific community with novel structures of macromolecular complexes at presently unmatched resolution. We have analyzed the reasons for which the objectives could not be reached beforehand and acted accordingly. The technological advances in cryo-EM and in cryo-ET are such that atomic resolutions can now be reached on frozen hydrated single molecules and specimen heterogeneity can be assessed and sorted using robust maximum-likelihood-based algorithms. In this respect a new development, Volta Phase plate [191, 192] (installed on the Titan Krios in IGBMC), provides robust signal in low-dose conditions in cryo-EM/ET facilitating the detection and separation of different conformational or structural states in our sample [153, 193]. The biochemical quality of the sample is crucial and A. Ben-Shem (in our laboratory) spent two years designing a purification protocol adapted for rare nuclear complexes that yields highly purified and stable complexes suitable for cryo-EM studies. As the stability of the samples limits our ability to reach high resolution we will take advantage of our recent experience with the SAGA complex [194] to propose a chemical approach, cross-linking with glutaraldehyde, to stabilize particular conformations of the complex.

In addition to the structural aspects, the collaborations amongst the project's Partners will allow to explore the promoter selection of the different type of native TFIID-containing PICs and their promoter-bound features genome-wide. All this information together will help to understand how native TFIID-containing PICs bind to distinct promoters and how they regulate transcription.

Tasks

Task1 – Characterization of reconstituted TFIID-containing complexes.

Rational and hypothesis: This project is organized around the central objective to understand the structure and function of the yeast TFIID complex. The main bottleneck for structural studies of endogenous large multi-subunit complexes is their low abundance in the cell. Large amounts of cells are therefore necessary to start structure determination. Lysing huge amounts of cells requires mechanical forces that produce heat and may dissociate labile complexes. Obtaining homogenous population of large and rare complexes in quantities suitable for cryo-EM or cryo-ET is a difficult task and currently one of the major limitations of their structural analysis. To circumvent this problem a novel purification scheme was established in our lab by Adam Ben-Shem. To understand how TFIID interacts with DNA and the GTFs in the PIC we plan to reconstitute early initiation complexes. In this task we aim to investigate the structure of TFIID bound to DNA, TFIIA or TFIIB.

Work program:

1.1 Protein purification

The objective of this part of the project is to obtain highly homogenous and pure TFIID complexes suitable for high-resolution structure determination by cryo-electron microscopy. A cornerstone of this project is the use of the yeast *Pichia pastoris* (Pp) to extract endogenous rare nuclear complexes. *P. pastoris* can be cultured to high cell density while maintaining an exponential growth rate. We found conditions to degrade the cell wall of Pp cells, which makes large-scale production from Pp nuclear extracts feasible for the first time. Whereas most previous structural studies of nuclear complexes made use of whole cell extracts produced by mechanical breaking of a large cell mass, we employ highly concentrated nuclear extracts, devoid of cytoplasmic contaminants and proteases, and do not expose the complexes to heat, mechanical forces or dilution. Purification from a concentrated nuclear extract requires a small number of steps and short incubation times. In addition, we employ affinity tags that can be rapidly eluted from affinity resins under mild conditions, maintaining relatively concentrated complexes at all times. TFIID bound DNA is eliminated by treating the sample with high-salt containing buffer. We reproducibly obtain sample concentrations of 1mg/ml. The novel purification method was tested with the transcriptional co-activator SAGA and we obtained a highly homogenous sample in a quantity suitable for structure determination with cryo-EM [194]. We tested the purification of TFIID by utilizing a streptavidin-binding peptide (SBP) affinity tag on the C-terminus of Taf2 (Figure 1). We will test the effect of tag placement on TFIID stability by introducing the SBP tag on other unique subunits of TFIID, like Taf4, Taf1 or Taf3 to ensure that we analyze the holo TFIID complex since the *in vivo* existence of TFIID subcomplexes was reported by L. Tora [145]. The structures determined in **Task 1.2** will facilitate the selection of subunits to be tagged.

TFIIA is a two subunit complex in yeast (TOA1 and TOA2), essential for a productive interaction of TFIID with promoter DNA. There are mainly two strategies for purification of TFIIA: 1) to overexpress and purify in *E. coli* as three polypeptides corresponding to the N-terminal as well as the C-terminal part of TOA1 and the entire TOA2 *[120]*; or 2) to overexpress TOA1 and TOA2 as fusion proteins [195]. We chose the first strategy, however in case of difficulties in the purification we will follow the second one.

TFIIB is a small protein (38 kDa) with high conservation among species and will be purified by overexpression in *E. coli* as described [18].

1.2 High resolution structure of reconstituted TFIID-containing complexes

The aim of this part of the project is to provide high-resolution structural information on functional TFIID complexes to describe the interaction interfaces of Tafs with different types of promoter DNAs, the TFIIA and TFIIB GTFs and with a well-defined activator. We aim to understand how cell signaling information is transmitted by TFIID to trigger PIC formation upon activator binding to the promoter DNA. We plan to obtain an atomic structure of the entire TFIID complex by combining single particle cryo-EM and existing X-ray structures of Taf subunits. We already obtained outstanding preliminary results by analyzing PpTFIID purified according to the new scheme (Figure 2). The intrinsic flexibility of TFIID was greatly reduced upon interacting with a TATA-box containing promoter thus yielding a preliminary 4.7 Å resolution cryo-EM map (Figure 2). However, one part of the structure, probably containing the key Taf1 and TBP subunits, is still poorly resolved. Additional interaction partners, such as TFIIB or transcriptional activators, will be incorporated to further stabilize TFIID. TFIIB was reported to interact directly with TFIID and to stabilize its interaction with the promoter DNA [196]. TFIIB also interacts with pol II [16], therefore the structure of TFIID-TFIIB could reveal the possible mode of TFIID incorporation into the PIC.

A second opportunity to stabilize the flexible TFIID lobe is to form an activation intermediate in which TFIID interacts with a promoter-bound activator. One of the best-studied systems is the Rap1 activator that regulates ribosomal protein expression and for which biochemical and genetic data demonstrate its direct binding to TFIID [42]. We anticipate that having TFIID bound to both the activator and to the proximal promoter through TBP will further stabilize the complex and allow us to improve resolution. Structural insights in such an activation complex would be of highest biological interest since to date no molecular model of an activator bound to TFIID is available.

The mode of TFIID interaction with TATA-less promoters is currently not known as well as the role of TBP in the absence of its cognate binding site. To understand how TFIID binds to promoter DNA we plan to use native TATA box containing or TATA-less promoter fragments. The best native target DNA sequences, or their consensus sequences, belonging to each category will be obtained from **Task 4.5**.

The XL-MS interaction maps will be performed for each functional TFIID complex (**Task 3) to** will help in positioning the different subunits into the EM density maps and to identify possible conformational changes.

Risk assessment: The stability of TFIID can vary upon tag usage. We plan to explore the effect of the tag position in this regard. In our experience DNA and other interacting partners greatly improves the complex stability. TBP establishes several interactions inside TFIID suggesting multiple binding sites and weaker interaction with the rest of the complex. TBP and TFIIA can be locked on the complex by the repressor/activator protein Rap1, which was reported to directly interact both with TFIID and with TFIIA [25].

Task2 – Characterization of endogenous TFIID-containing transcription complexes

Rational and hypothesis: The preinitiation complexes reconstituted *in vitro* from recombinant proteins do not reflect the whole composition of an *in vivo* system. Besides the role of TFIID in PIC formation is poorly understood. The objective of this task is to provide information about the structure and composition of the native

TFIID-containing promoter-bound transcription complexes and how TFIID participates in the activated transcription initiation using biochemistry, mass-spectrometry, cryo-EM and cryo-ET.

2.1 Purification of TFIID-containing complexes bound to DNA

The objective of this part of the project is to obtain TFIID-containing transcription complexes suitable for high-resolution structure determination by cryo-ET and cryo-EM.

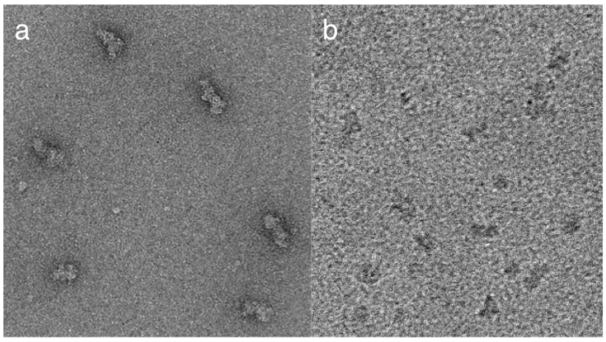


Figure 4 Purified yeast Preinitiation complexes in negative stain (**a**) and in frozen hydrated state (**b**).

To purify the TFIID-containing native promoter-bound transcription complexes we will use two successive purification method steps, where in addition to TFIID we will tag other members of the PIC with a second affinity tag. The feasibility of the method applied for TFIID-containing PICs was tested using SBP-tag on Taf2 (TFIID) and His tag on Rpb2 (Pol II), which identified DNA-bound supramolecular assemblies. Mass-spectrometry confirmed the presence of all previously published stable components of the PIC together with the Mediator and SAGA. Interestingly we also detected the presence of chromatin remodeling complexes and transcription activators opening the possibility to investigate the architecture of and interactions involving transcription regulatory machines including chromatin regulatory complexes. This preliminary result suggests that we are able to purify endogenous TFIID and Pol II containing complexes, which were not studied before. Negative-stain and first cryo-EM showed that the specimen will be suitable for structure determination (Figure 4). In this Task, we will tag additional subunits in an effort to further optimize the system to obtain TFIID-associated transcription complexes of high quality and quantity sufficient for cryo-electron microscopy and cryo-electron tomography. The DNA fragments, bound by the purified TFIID-containing complexes, will be preserved by avoiding the high-salt treatment and will be purified for subsequent sequence analysis in **Task4** to define the promoter selectivity of different TFIID-containing complexes and correlate them to transcriptome analysis and genome-wide active promoter mapping.

2.2 Subunit composition of the native PICs by mass-spectrometry

The subunit composition of the purified TFIID-containing promoter-bound transcription complexes will be analyzed using mass-spectrometry. This step is essential to assess the quality of the purification and gives information about the composition of the native *Pichia pastoris* TFIID-containing complexes. Heterogeneous samples will be further separated by gradient centrifugation or native electrophoresis methods prior of mass-spectrometry to determine their protein composition separately. The use of tags on different components of TFIID along with the tag on Pol II (as well as a potential third tag on a GTF; see below) in the tandem-affinity purifications could likewise reveal heterogeneity in native preinitiation complexes. We will test this hypothesis by defining the protein composition of the isolated double, or triple purified PICs by mass-spectrometry. Note that this sub-fractionation is not required for electron microscopy analysis since subpopulations will be separated *in silico* by image processing.

2.3 Structure of the native TFIID-containing complexes

The composition of the TFIID-containing transcription complexes may vary from gene to gene or according to PIC assembly step. To deal with this heterogeneity cryo-electron tomography will be used to reconstruct 3-D models of individual complexes and will be combined with sub-tomogram averaging followed by classification to improve resolution of homogeneous classes of PICs. The resulting models will be refined by single particle cryo-EM. The large components of the PIC will be identified by their shape and confirmed by mass spectrometry. In combination with XL-MS, the cryo-EM maps will allow us to reveal the structural organization and heterogeneity of each complex type. Both cryo-electron microscopy and tomography have made major progresses in the recent years. Single particle cryo-EM produces structures based on the hypothesis that the imaged objects are identical or similar and that they differ by the angle of view. Atomic resolution can be reached by cryo-EM, however the technique has its limitation to sort structural heterogeneities. On the other side cryo-ET created structures of the individual objects by physically tilting the specimen and thereby obtaining different views from the same entity and a greater degree of heterogeneity can thus be sorted. Recently introduced maximum-likelihood method based techniques facilitate the sorting of structural heterogeneity in 2-D image datasets and will also be used to separate the various types of TFIID-containing complexes. New direct electron detectors and image processing algorithms can now provide unprecedented resolution, however all these methods rely on signal detection in the noisy cryo-EM or even noisier cryo-ET images. The signal-to-noise ratio in the images can be drastically improved by using a lately developed instrument, the Volta phase plate [197], available on site. This device, by introducing a phase shift in the electron beam, greatly enhances the contrast in the images therefore the heterogeneity detection and facilitates image alignment.

Risk assessment: The tandem-affinity purification of native PICs worked well in the preliminary experiments. We plan to utilize other tags besides the His tag to increase the purification yield. A large spectrum of tags can be incorporated into various components of TFIID and other PIC components, like HA, FLAG, choline-binding domain or protein A tag. We plan to explore the possibility of using different tagged PIC subunits for the purification and compare the results using mass-spectrometry and electron microscopy. The native TFIID-containing complexes are presumably fragile and we plan to stabilize them by chemical crosslinking with glutaraldehyde using the GraFix method [99]. If necessary, we will also explore the possibility of briefly treating cells with a low concentration of formaldehyde as a way to stabilize PICs prior to purification. This approach has been used successfully to stabilize *in vivo* relevant interactions prior to purification of a number of macromolecular complexes [198]. The possible variations in the composition of the native PICs can introduce difficulties to the project. We plan to improve the detection of the structural heterogeneity by utilizing the Volta phase plate. At 10 Å resolution the forms of the large components are readily detectable and the combination of XL-MS and integrative modeling can provide the position of all components. We now have a palette of structures of the GTFs accessible along with the Mediator and SAGA, which can be used to identify the different complexes in the cryo-EM maps of native TFIID-containing complexes.

Task3 – The subunit interaction landscape of promoter DNA associated TFIID complexes

Rational and hypothesis: Numerous studies have been devoted to studying the composition and architecture of complexes involved in transcription initiation. Several structures of reconstituted PIC subassemblies have been solved, however the exact interaction network of a native PIC is still unknown. We aim to define this interaction map of *in vivo* TFIID-containing complexes by the help of crosslinking coupled with mass-spectrometry (XL-MS). We will build upon mapping the subunit organization of TFIID to define the interaction landscape of early TFIID-containing intermediates in PIC formation.

Work program:

3.1 Interaction map of TFIID and its complexes by XL-MS

The goal of this Aim is to map the subunit organization of TFIID in association with TATA box containing or TATA less promoter DNA fragments, with either TFIIA and/or TFIIB by XL-MS.

As described above, we have already collaborated with J. Ranish successfully to define the subunit interaction landscape of *Pp*TFIID by XL-MS (Figure 3). In this aim will we build upon these studies by applying XL-MS to analyze the TFIID-IIA-IIB promoter complex. Recent studies in the human system suggest that TFIID undergoes a conformational rearrangement upon association with promoter DNA and TFIIA (40). By comparing crosslinks identified with free TFIID to those identified with the DNA bound TFIID with TFIIA or TFIIB, we may be able to identify changes in the TFIID crosslink maps that could be due to conformational changes in TFIID. We have large quantities of purified TFIID, -IIA, and –IIB which will be used to assemble the individual complexes on promoter DNA (see **Task 1.2**). DNA bound TFIID-containing complexes will be isolated by glycerol gradient sedimentation, and the purified complexes will then be crosslinked with the homo-bi-functional amine-reactive crosslinking reagent BS3 (Thermo-Scientific) and analyzed by high resolution mass spectrometry using protocols that are well established in the laboratory of J. Ranish [27, 199-201]. Identification of BS3 crosslinked peptides will be performed using the Nexus2.0 (developed by J. Ranish, unpublished) and pLINK [202] database search algorithms which also estimate false discovery rates. All identified crosslinks will be verified by manual inspection of the data, and at least two independent experiments will be performed with each complex. The crosslinks will then be mapped onto the sequences of the proteins in the complex, including domain information, to create linkage maps which will be used to infer protein-protein and domain-domain interactions within the complexes. By comparing the linkage maps for the different complexes, we may identify differences that could be due to changes in the architecture of the complexes. For example, we might detect changes in the crosslinking patterns for DNA bound TFIID vs. unbound TFIID, or for TFIID bound to TATA vs TATA less DNA. We will also attempt to generate molecular models of the complexes by integrating the crosslink data with EM maps (Figure 2 and **Task 1.2**).

3.2 Interaction map of the native PICs by XL-MS

Thus far, little information is known about the subunit arrangement of PICs *in vivo*. The objective of this task is to provide information about the interaction landscape of native TFIID-containing promoter-bound transcription complexes and how TFIID participates in transcription initiation using XL-MS. TFIID-containing transcription complexes isolated by the multi-step affinity purification method described in **Task 1** will be subjected to BS3 crosslinking, mass analysis and database searching as described in **Task 3.1**. A possible issue for this analysis could be the quantity of the purified sample. We typically require ~40 pmols of material to identify large numbers of crosslinked peptides. If necessary we can scale up the amount of cells used to prepare nuclear extracts to obtain sufficient material for successful XL-MS analysis. Another potential issue is heterogeneity of the samples due for example to isolation of partial complexes and/or PICs assembled on different promoters. This issue should be mitigated by the use of a multi-step affinity purification scheme using tags on different components of the PIC. Even if the samples are heterogeneous, they will be enriched for TFIID-containing transcription complexes. One of the advantages of XL-MS for the study protein interactions is its ability to reveal protein-protein interactions (PPIs) in complex samples. We expect that PPIs that are present in most of the complexes will be the most likely to be identified by XL-MS, while those that are present in some of the complexes will be identified at a lower frequency. Our database search algorithms can confidently identify crosslinked peptides using databases composed of ~100 proteins. Since ~100 PIC components are known, we should be able to confidently identify crosslinked peptides in these samples using this approach. The results should provide, for the first time, important information, about the interaction landscape of native TFIID-containing transcription complexes. This information will be used to create protein-protein and domain-domain linkage maps for TFIID-containing complexes as described in **Task 2.1**. It will also be used to complement the ET and EM maps that will be generated in **Task 2.3**.

Risk assessment: We anticipate no problem with Task 3.1 as we have already generated dense crosslinking maps for TFIID. We will also consider using crosslinking reagents with different reactivities and physico-chemical characteristics such EDC or DSS, respectively which can provide complementary distance restraints to BS3. EDC links amines to carboxylic acids and DSS is an amine reactive crosslinking, which is more hydrophobic than BS3. Task 3.2 is more challenging due to the increased complexity of the samples and the potential challenges of isolating sufficient material for XL-MS analysis. Nonetheless we have the ability to scale the purification to obtain the required amount of sample and our XL-MS approach can confidently identify crosslinked peptides derived from samples containing ~100 proteins. J. Ranish also has developed XL-MS approaches that take advantage of MS labile crosslinkers [203]. The special properties of MS labile crosslinkers permit confident identification of crosslinked peptides by searching whole proteome databases. These crosslinkers can be used to complement the BS3-based experiments and/or to overcome potential sample complexity issues. If necessary we can treat cells with a low concentration of formaldehyde for a short time period, in order to stabilize PICs prior to affinity purification. Formaldehyde treatment has been used successfully to stabilize complexes prior to BS3 crosslinking [204].

Task4 – Promoter DNA analysis, transcriptome analysis and genome-wide active promoter mapping

Rational and hypothesis: TFIID is one of the first GTFs to recognize promoters and thus, triggering transcription initiation. To define the transcriptionally active promoter-bound TFIID-containing PICs, first L. Tora will sequence the bound DNA of the native TFIID-containing complexes (see **Task 2**); second L. Tora will map these binding sites to the *Pichia pastoris* (*Pp*) genome; third L. Tora will carry out chromatin immunoprecipitation (ChIP)-coupled sequencing (ChIP-seq) to map the binding of TFIID and TFIID-containing PICs on the *Pp* genome *in vivo*; third we will determine which of the *Pp* promoters were active by using nascent RNA seq; fifth, we will carry out bioinformatics analyses by comparing the data sets to define which transcriptionally active endogenous promoters were bound by TFIID/PICs, and to investigate whether DNA sequence binding in active promoters by TFIID/PICs can be categorized in the light of the structural data obtained above. Using anti-RNA Pol II as a read-out for PIC binding and/or transcription itself, L. Tora's lab has uncovered several important regulatory mechanisms targeting Pol II transcription on a global genome-wide scale [205-207]. To analyze direct read-outs for transcription, L. Tora's lab has used very recently the quantification of newly-synthesized mRNA levels genome-wide in yeast by 4-thiouridine (4SU)-based RNA tagging [28]. We will use this combined know-how to achieve the following steps:

Work program:

4.1 DNA interactome of endogenous TFIID-containing complexes

We will purify the bound DNA fragments from the endogenous TFIID/PIC-containing complexes obtained in **Task2**. An adapter will be ligated to the DNA ends and they will be amplified by PCR and sequenced by high-throughput sequencing. To define promoter occupancy of native TFIID-containing PIC complexes we will correlate the obtained sequences with transcriptome analysis (**Task 4.2**) and active promoter mapping (**Task 4.3**) along with combination of tags used in the tandem-affinity purification (**Task 2.1**).

4.2 Mapping the presence of TFIID/PICs genome-wide

To map TFIID, GTF and PIC occupancy in Pp genome-wide, we will carry out chromatin immunoprecipitation (ChIP) coupled to high throughput sequencing (seq) analyses using mAbs or affinity resins that recognize either SBP tagged subunits of TFIID, i.e. Taf2 (see above), or other tagged Tafs, or subunits of the different GTFs (see Task 2.3) together with antibodies recognizing the non-phosphorylated or different phosphorylated forms of the repeats present in C-terminal domain (CTD) of the largest subunit of Pol II (Rpb1). If the commercially available antibodies against the Ser2 (representative of elongating Pol II) or Ser5 (representative of initiating Pol II) phosphorylated forms of the Pp CTD will not work in ChIP, or if other Pp-specific antibodies will be required for ChIP, L. Tora will raise antibodies at the antibody facility on site. Cells will be fixed with formaldehyde and ChIP carried out, purified DNA will be sonicated and deep sequenced (at the IGBMC high throughput sequencing platform). L. Tora has used this technology already in yeast [10]. Specific Taf/TFIID, GTF or Pol II bound sequence-reads will be mapped to the *Pp* genome, and unique reads will be considered for further analyses. For the comparative ChIP-seq analyses, *Schizosaccharomyces pombe* cells will be mixed with the *Pp* cells, as spike in controls. Next, TFIID, GTF or Pol II density profiles on the coding regions of all refseq genes will be calculated for all datasets by using the seqMINER tool developed in L. Tora's team [208]. Different ChIP-seq data sets obtained for different TFIID, GTF and Pol II forms will be calculated and represented by k-means clustering. These combined comparisons will define the presence of TFIID-, GTF- and/or Pol II-containing PICs at the distinct promoters of *Pp* cells genome-wide. Moreover, pathway analyses for these potentially different TFIID-containing PIC-classes will be carried out using several bioinformatics tools. These results will also be compared to results obtained in **Task 4.1.** (see also Task 4.4).

4.3 Analyzing active gene promoters genome-wide

Genome-wide RNA profiling technologies greatly facilitate the global analysis of gene expression. However, such technologies often do not give a direct information on RNA transcription but rather a mixed read-out on transcription/synthesis and RNA decay. To overcome such limitation and to get a direct read-out of transcription, metabolic labeling of newly synthesized RNA with 4-thiouridine (4sU) combined with genome-wide RNA profiling will be used to measure directly newly synthetized RNA transcription (and decay). L. Tora has used this technology recently in several publications [28, 30]. To map transcriptionally active gene promoters genome-wide, we will use 4sU metabolic labeling of newly synthesized RNA, a biotinylation-based purification coupled RNA sequencing technology that detects 4-thiouridine (4sU) incorporation in newly synthetized RNA species. Cells will be exposed for 6-10 min to 4sU, at the end of 4sU exposure, cells will be harvested, total cellular RNA will be prepared, and newly transcribed RNA 4sU-containing RNA will be biothinylated and purified on streptavidin beads. Newly transcribed RNA will be subjected to RNA-seq analyses (at the IGBMC deep sequencing platform). In collaboration with the IGBMC bioinformatics platform newly transcribed transcripts will be mapped to Pp genome and active gene promoters determined.

4.4 Bioinformatics analyses

Data sets will be bioinformatically compared containing either active promoters (obtained in Task 4.3), or the TFIID, GTF and Pol II occupancy profiles (obtained Task 4.2), or the DNA interactome of endogenous TFIID-containing complexes (obtained in Task 2 and sequenced in Task 4.1). L. Tora’s lab has developed bioinformatics tools (called seqMiner) to carry out such analyses [208]. These combined analyses will allow the identification of active PICs in *Pichia pastoris*. Moreover, these combined analyses may reveal for the first-time potential variations in the composition of PICs during transcription initiation on different subset of promoters. In addition, we plan to investigate whether different PIC compositions may be determined by distinct core-promoter sequence elements. Active TFIID-containing promoters will be classified on the basis of their sequence preference and their consensus sequences will be defined. The tagging and consequent ChIP-sequencing of several Taf subunits (present in different lobes of TFIID) will also allow us to determine whether TFIID always acts as a holo complex, or whether partial Taf/TFIID assemblies may also have functional role *in vivo*. The combined and related information obtained by these bioinformatics analyses can then be further incorporated and utilized in the structural **Tasks** described above to ameliorate and to better characterize the obtained structures and their relevance.

Risk assessment: As the above described methodologies and related analyses are routinely carried out in L. Tora’s laboratory we do not foresee major hurdles when carrying out this part of the project.

REFERENCES

1. Papai, G., et al., Structure, **17**, 363-73, (2009).

2. Grob, P., et al., Structure, **14**, 511-20, (2006).

3. Leurent, C., et al., Embo J, **21**, 3424-33, (2002).

4. Sanders, S. L., et al., Mol Cell Biol, **22**, 6000-13, (2002).

5. Hahn, S., Nat Struct Mol Biol, **11**, 394-403, (2004).

6. Thomas, M. C., et al., Crit Rev Biochem Mol Biol, **41**, 105-78, (2006).

7. Grunberg, S., et al., Trends Biochem Sci, **38**, 603-11, (2013).

8. Roeder, R. G., Trends in biochemical sciences, **21**, 327-35, (1996).

9. Kim, B., et al., Proc Natl Acad Sci U S A, **104**, 16068-73, (2007).

10. Bonnet, J., et al., Genes Dev, **28**, 1999-2012, (2014).

11. Flanagan, P. M., et al., Nature, **350**, 436-8, (1991).

12. Ohler, U., et al., Genome Biol, **3**, RESEARCH0087, (2002).

13. Yang, C., et al., Gene, **389**, 52-65, (2007).

14. Rhee, H. S., et al., Nature, **483**, 295-301, (2012).

15. Forget, D., et al., Proc Natl Acad Sci U S A, **94**, 7150-5, (1997).

16. Kostrewa, D., et al., Nature, **462**, 323-30, (2009).

17. Liu, X., et al., Science, **327**, 206-9, (2010).

18. Sainsbury, S., et al., Nature, **493**, 437-40, (2013).

19. Kang, J. J., et al., Mol Cell Biol, **15**, 1234-43, (1995).

20. Ranish, J. A., et al., Genes Dev, **13**, 49-63, (1999).

21. Rani, P. G., et al., Mol Cell Biol, **24**, 1709-20, (2004).

22. Lescure, A., et al., Embo j, **13**, 1166-75, (1994).

23. Wieczorek, E., et al., Nature, **393**, 187-91, (1998).

24. Brou, C., et al., Embo j, **12**, 489-99, (1993).

25. Papai, G., et al., Nature, **465**, 956-60, (2010).

26. Bieniossek, C., et al., Nature, **493**, 699-702, (2013).

27. Luo, J., et al., Mol Cell, **59**, 794-806, (2015).

28. Baptista, T., et al., Mol Cell, **68**, 130-43 e5, (2017).

29. Gupta, K., et al., Elife, **6**, (2017).

30. Warfield, L., et al., Mol Cell, **68**, 118-29 e5, (2017).

31. Burley, S. K., et al., Annu Rev Biochem, **65**, 769-99, (1996).

32. Green, M. R., Trends in biochemical sciences, **25**, 59-63, (2000).

33. Mohan, W. S., Jr., et al., Mol Cell Biol, **23**, 4307-18, (2003).

34. Muller, F., et al., Curr Opin Genet Dev, **20**, 533-40, (2010).

35. Vermeulen, M., et al., Cell, **131**, 58-69, (2007).

36. Hampsey, M., et al., Curr Opin Genet Dev, **9**, 132-9, (1999).

37. Juven-Gershon, T., et al., Curr Opin Cell Biol, **20**, 253-9, (2008).

38. Dynlacht, B. D., et al., Cell, **66**, 563-76, (1991).

39. Gill, G., et al., Proceedings of the National Academy of Sciences of the United States of America, **91**, 192-6, (1994).

40. Asahara, H., et al., Molecular and cellular biology, **21**, 7892-900, (2001).

41. Liu, X., et al., Molecular and cellular biology, **13**, 3291-300, (1993).

42. Garbett, K. A., et al., Mol Cell Biol, **27**, 297-311, (2007).

43. Louder, R. K., et al., Nature, **531**, 604-9, (2016).

44. Lee, D. H., et al., Mol Cell Biol, **25**, 9674-86, (2005).

45. Martinez, E., et al., Embo J, **13**, 3115-26, (1994).

46. Verrijzer, C. P., et al., Cell, **81**, 1115-25, (1995).

47. Burke, T. W., et al., Genes Dev, **11**, 3020-31, (1997).

48. Johannessen, M., et al., J Gen Virol, **84**, 1887-97, (2003).

49. Lavigne, A. C., et al., J Biol Chem, **271**, 19774-80, (1996).

50. Dubrovskaya, V., et al., Embo j, **15**, 3702-12, (1996).

51. Trowitzsch, S., et al., Nat Commun, **6**, 6011, (2015).

52. Bai, Y., et al., Mol Cell Biol, **17**, 3081-93, (1997).

53. Bagby, S., et al., FEBS Lett, **468**, 149-54, (2000).

54. Kokubo, T., et al., Molecular and cellular biology, **18**, 1003-12, (1998).

55. Hoffmann, A., et al., Nature, **380**, 356-9., (1996).

56. Xie, X., et al., Nature, **380**, 316-22, (1996).

57. Birck, C., et al., Cell, **94**, 239-49, (1998).

58. Werten, S., et al., J Biol Chem, **277**, 45502-9, (2002).

59. Gangloff, Y. G., et al., Molecular and cellular biology, **21**, 1841-53, (2001).

60. Gangloff, Y. G., et al., Mol Cell Biol, **20**, 340-51, (2000).

61. Jacobson, R. H., et al., Science, **288**, 1422-5, (2000).

62. Gangloff, Y. G., et al., Mol Cell Biol, **21**, 5109-21, (2001).

63. Feigerle, J. T., et al., J Biol Chem, **291**, 22721-40, (2016).

64. Cianfrocco, M. A., et al., Cell, **152**, 120-31, (2013).

65. van Holde, K., (1988).

66. Luger, K., et al., Nature, **389**, 251-60, (1997).

67. Hamiche, A., et al., J Mol Biol, **257**, 30-42, (1996).

68. Bednar, J., et al., Proc Natl Acad Sci U S A, **95**, 14173-8, (1998).

69. Woodcock, C. L., et al., Chromosome Res, **14**, 17-25, (2006).

70. Thoma, F., et al., J Cell Biol, **83**, 403-27, (1979).

71. Bednar, J., et al., J Cell Biol, **131**, 1365-76, (1995).

72. Tremethick, D. J., Cell, **128**, 651-4, (2007).

73. Maresca, T. J., et al., J Cell Biol, **169**, 859-69, (2005).

74. Izzo, A., et al., Biol Chem, **389**, 333-43, (2008).

75. Allan, J., et al., Nature, **288**, 675-9, (1980).

76. Hendzel, M. J., et al., J Biol Chem, **279**, 20028-34, (2004).

77. Syed, S. H., et al., Proc Natl Acad Sci U S A, **107**, 9620-5, (2010).

78. Simpson, R. T., Biochemistry, **17**, 5524-31, (1978).

79. Allan, J., et al., J Mol Biol, **187**, 591-601, (1986).

80. Fang, H., et al., Nucleic Acids Res, **40**, 1475-84, (2012).

81. Ramakrishnan, V., et al., Nature, **362**, 219-23, (1993).

82. Staynov, D. Z., et al., EMBO J, **7**, 3685-91, (1988).

83. Zhou, Y. B., et al., Nature, **395**, 402-5, (1998).

84. Pruss, D., et al., Science, **274**, 614-7, (1996).

85. Brown, D. T., et al., Nat Struct Mol Biol, **13**, 250-5, (2006).

86. Zhou, B. R., et al., Mol Cell, **59**, 628-38, (2015).

87. Zhou, B. R., et al., Proc Natl Acad Sci U S A, **110**, 19390-5, (2013).

88. Zhou, B. R., et al., J Mol Biol, (2016).

89. Song, F., et al., Science, **344**, 376-80, (2014).

90. Fan, L., et al., Proc Natl Acad Sci U S A, **103**, 8384-9, (2006).

91. Leurent, C., et al., Embo J, **23**, 719-27, (2004).

92. Brand, M., et al., Science, **286**, 2151-3, (1999).

93. Andel, F., 3rd, et al., Science, **286**, 2153-6, (1999).

94. Callebaut, I., et al., BMC Genomics, **6**, 100, (2005).

95. Sanders, S. L., et al., J Biol Chem, **275**, 13895-900, (2000).

96. Rigaut, G., et al., Nat Biotechnol, **17**, 1030-2, (1999).

97. Kaufmann, J., et al., Genes Dev, **10**, 873-86, (1996).

98. Golas, M. M., et al., Mol Cell, **17**, 869-83, (2005).

99. Kastner, B., et al., Nat Methods, **5**, 53-5, (2008).

100. Thunnissen, M. M., et al., Nat Struct Biol, **8**, 131-5, (2001).

101. Leschziner, A. E., et al., Annu Rev Biophys Biomol Struct, **36**, 43-62, (2007).

102. Liu, W. L., et al., Mol Cell, **29**, 81-91, (2008).

103. Verrijzer, C. P., et al., Science, **264**, 933-41, (1994).

104. Banik, U., et al., J Biol Chem, **276**, 49100-9, (2001).

105. Chalkley, G. E., et al., Embo J, **18**, 4835-45, (1999).

106. Oelgeschlager, T., et al., Nature, **382**, 735-8, (1996).

107. Wright, K. J., et al., Proc Natl Acad Sci U S A, **103**, 12347-52, (2006).

108. Singh, M. V., et al., Mol Cell Biol, **24**, 4929-42, (2004).

109. Tasto, J. J., et al., Yeast, **18**, 657-62, (2001).

110. Kremer, J. R., et al., J Struct Biol, **116**, 71-6, (1996).

111. Pettersen, E. F., et al., J Comput Chem, **25**, 1605-12, (2004).

112. van Heel, M., et al., J Struct Biol, **116**, 17-24, (1996).

113. Frank, J., et al., J Struct Biol, **116**, 190-9, (1996).

114. van Heel, M., et al., J Struct Biol, **151**, 250-62, (2005).

115. Elmlund, H., et al., Structure, **17**, 1442-52, (2009).

116. Sussel, L., et al., Proc Natl Acad Sci U S A, **88**, 7749-53, (1991).

117. Lieb, J. D., et al., Nat Genet, **28**, 327-34, (2001).

118. Liu, W. L., et al., Genes Dev, **23**, 1510-21, (2009).

119. Layer, J. H., et al., Submitted.

120. Tan, S., et al., Nature, **381**, 127-51, (1996).

121. Geiger, J. H., et al., Science, **272**, 830-6, (1996).

122. Mencia, M., et al., Mol Cell, **9**, 823-33, (2002).

123. Simonetti, A., et al., Nature, **455**, 416-20, (2008).

124. Konig, P., et al., Cell, **85**, 125-36, (1996).

125. Morse, R. H., Trends Genet, **16**, 51-3, (2000).

126. Bleichenbacher, M., et al., J Mol Biol, **332**, 783-93, (2003).

127. Wang, W., et al., Genes Dev, **6**, 1716-27, (1992).

128. Lieberman, P. M., et al., Mol Cell Biol, **17**, 6624-32, (1997).

129. Ozer, J., et al., J Biol Chem, **271**, 11182-90, (1996).

130. Shykind, B. M., et al., Genes Dev, **9**, 1354-65, (1995).

131. Kokubo, T., et al., Mol Cell Biol, **18**, 1003-12, (1998).

132. Rathjen, J., et al., Nucleic Acids Res, **18**, 3219-25, (1990).

133. Ludtke, S. J., et al., J Struct Biol, **128**, 82-97, (1999).

134. Tang, G., et al., J Struct Biol, **157**, 38-46, (2007).

135. Heymann, J. B., J Struct Biol, **133**, 156-69, (2001).

136. Cler, E., et al., Cell Mol Life Sci, **66**, 2123-34, (2009).

137. Papai, G., et al., Current opinion in genetics & development, **21**, 219-24, (2011).

138. Bhattacharya, S., et al., Proc Natl Acad Sci U S A, **104**, 1189-94, (2007).

139. Romier, C., et al., J Mol Biol, **368**, 1292-306, (2007).

140. Scheer, E., et al., J Biol Chem, **287**, 27580-92, (2012).

141. Fitzgerald, D. J., et al., Structure, **15**, 275-9, (2007).

142. Wang, X., et al., Proc Natl Acad Sci U S A, **104**, 7839-44, (2007).

143. Selleck, W., et al., Nature structural biology, **8**, 695-700, (2001).

144. Soutoglou, E., et al., Mol Cell Biol, **25**, 4092-104, (2005).

145. Demeny, M. A., et al., PLoS One, **2**, e316, (2007).

146. Timmers, H. T., et al., Trends Biochem Sci, **30**, 7-10, (2005).

147. Mengus, G., et al., EMBO J, **24**, 2753-67, (2005).

148. Bell, B., et al., Mol Cell, **8**, 591-600, (2001).

149. Chen, Z., et al., J Biol Chem, **278**, 35172-83, (2003).

150. Bhattacharya, S., et al., Proc Natl Acad Sci U S A, **111**, 9103-8, (2014).

151. Thuault, S., et al., J Biol Chem, **277**, 45510-7, (2002).

152. Liu, D., et al., Cell, **94**, 573-83, (1998).

153. Danev, R., et al., Elife, **5**, (2016).

154. Wang, H., et al., Cell Res, **24**, 1433-44, (2014).

155. Shao, H., et al., Mol Cell Biol, **25**, 206-19, (2005).

156. Gazit, K., et al., J Biol Chem, **284**, 26286-96, (2009).

157. Mattiroli, F., et al., Science, **357**, 609-12, (2017).

158. Hahn, S., et al., Genetics, **189**, 705-36, (2011).

159. Widom, J., Proc Natl Acad Sci U S A, **89**, 1095-9, (1992).

160. Joo, Y. J., et al., Genes Dev, **31**, 2162-74, (2017).

161. Layer, J. H., et al., J Biol Chem, **288**, 23273-94, (2013).

162. Layer, J. H., et al., The Journal of biological chemistry, **285**, 15489-99, (2010).

163. Xu, Y., et al., Cancer Cell, **33**, 13-28 e8, (2018).

164. Riffle, M., et al., Journal of proteome research, **15**, 2863-70, (2016).

165. Ranish, J. A., et al., Science, **255**, 1127-9, (1992).

166. Mastronarde, D. N., J Struct Biol, **152**, 36-51, (2005).

167. Zheng, S. Q., et al., Nat Methods, **14**, 331-2, (2017).

168. Zhang, K., J Struct Biol, **193**, 1-12, (2016).

169. Scheres, S. H., J Mol Biol, **415**, 406-18, (2012).

170. Kucukelbir, A., et al., Nat Methods, **11**, 63-5, (2014).

171. Yang, J., et al., Nucleic Acids Res, **43**, W174-81, (2015).

172. Garzon, J. I., et al., Bioinformatics, **23**, 427-33, (2007).

173. Emsley, P., et al., Acta crystallographica. Section D, Biological crystallography, **66**, 486-501, (2010).

174. Kelley, L. A., et al., Nat Protoc, **4**, 363-71, (2009).

175. Terwilliger, T. C., Acta crystallographica. Section D, Biological crystallography, **66**, 268-75, (2010).

176. Goddard, T. D., et al., J Struct Biol, **157**, 281-7, (2007).

177. Goddard, T. D., et al., Protein Sci, **27**, 14-25, (2018).

178. Lowary, P. T., et al., J Mol Biol, **276**, 19-42, (1998).

179. Chua, E. Y., et al., Nucleic Acids Res, **40**, 6338-52, (2012).

180. Thomas, J. O., et al., EMBO J, **5**, 3531-7, (1986).

181. Mirzabekov, A. D., et al., J Mol Biol, **211**, 479-91, (1990).

182. Buckle, R. S., et al., J Mol Biol, **223**, 651-9, (1992).

183. de Vries, S. J., et al., Nat Protoc, **5**, 883-97, (2010).

184. Trott, O., et al., Journal of computational chemistry, **31**, 455-61, (2010).

185. Shukla, M. S., et al., Nucleic Acids Res, **39**, 2559-70, (2011).

186. Lone, I. N., et al., PLoS Genet, **9**, e1003830, (2013).

187. Oberg, C., et al., J Mol Biol, **419**, 183-97, (2012).

188. Wisniewski, J. R., et al., Mol Cell Proteomics, **6**, 72-87, (2007).

189. Wisniewski, J. R., et al., Nucleic Acids Res, **36**, 570-7, (2008).

190. Christophorou, M. A., et al., Nature, **507**, 104-8, (2014).

191. Danev, R., et al., Proc Natl Acad Sci U S A, **111**, 15635-40, (2014).

192. von Loeffelholz, O., et al., J Struct Biol, (2018).

193. Khoshouei, M., et al., J Struct Biol, **197**, 94-101, (2017).

194. Sharov, G., et al., Nat Commun, **8**, 1556, (2017).

195. Adachi, N., et al., Protein expression and purification, **133**, 50-6, (2017).

196. Hisatake, K., et al., Nature, **363**, 744-7, (1993).

197. Fukuda, Y., et al., J Struct Biol, **190**, 143-54, (2015).

198. Kaake, R. M., et al., Journal of proteome research, **9**, 2016-29, (2010).

199. McDermott, S. M., et al., Proc Natl Acad Sci U S A, **113**, E6476-e85, (2016).

200. Han, Y., et al., EMBO J, **33**, 2534-46, (2014).

201. Knutson, B. A., et al., Nat Struct Mol Biol, **21**, 810-6, (2014).

202. Yang, B., et al., Nat Methods, **9**, 904-6, (2012).

203. Luo, J., et al., Mol Cell Proteomics, **11**, M111.008318, (2012).

204. Robinson, P. J., et al., Cell, **166**, 1411-22.e16, (2016).

205. Gyenis, A., et al., PLoS Genet, **10**, e1004483, (2014).

206. Ravens, S., et al., Elife, **3**, (2014).

207. Ravens, S., et al., Epigenetics Chromatin, **8**, 45, (2015).

208. Ye, T., et al., Nucleic Acids Res, **39**, e35, (2011).