Dissection de TFIID, un facteur de transcription général humain : Études structurales etfonctionnelles des sous-ensembles du TFIID human

Kapil Gupta

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Biologie Structurale et Nanobiologie

Arrêté ministériel : 25 mai 2016

Présentée par

Kapil GUPTA

Thèse dirigée par **Imre BERGER (CSV)**, ,

préparée au sein du **Laboratoire laboratoire européen de biologie moléculaire**
dans **l'École Doctorale Chimie et Sciences du Vivant**

Dissection de TFIID, un Facteur de Transcription Général : Études Structurales et Fonctionnelles des Sous-ensembles du TFIID Humain

Dissecting General Transcription Factor TFIID: Structural and Functional Studies of Human TFIID Subassemblies

Thèse soutenue publiquement le **24 septembre 2015**,
devant le jury composé de :

Monsieur IMRE BERGER
PROFESSEUR, EMBL GRENOBLE, Directeur de thèse
Monsieur MARTIN BLACKLEDGE
DIRECTEUR DE RECHERCHE, CEA GRENOBLE, Président
Monsieur MICHAEL PLEVIN
MAITRE DE CONFERENCES, UNIVERSITE D'YORK - ROYAUME-UNI, Rapporteur
Monsieur STEPHANE THORE
CHARGE DE RECHERCHE, INSERM BORDEAUX, Rapporteur

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First of all, I would like thank my supervisor Prof. Imre Berger for giving me opportunity to work on these challenging projects and the guidance he provided during my PhD work.

Also, I would like to thank my Thesis Advising Committee members Prof. Christiane Berger-Schaffitzel, Dr. Andrew McCarthy, Dr. Christoph Mueller and Dr. Carlo Petosa for all the helpful discussions and suggestions during my annual TAC meetings.

I am grateful to my jury members Prof. Stephane Thore, Dr. Michael Plevin, Dr. Martin Blackledge for their time in reviewing my PhD thesis.

A number of collaborators contributed to this work. I would like to thank Dr. Malene Ringkjøbing Jensen in the Blackledge group (IBS, Grenoble) for NMR data collection, analysis and structure determination. Prof. Dame Carol Robinson and Dr. Ima Obong-Ebong (University of Oxford) carried out native MS studies, Juan Zou and Dr. Juri Rappsilber (University of Edinburgh) performed CLMS studies and Dr. Aleksandra Watson and Prof. Ernest Laue (University of Cambridge) are acknowledged for help in structural modelling of TAF11/TAF13/TBP complex. I would also like to thank Dr. Christophe Romier (IGBMC Straßbourg), as well as Dr. Christoph Bieniossek, Dr. Eaazhisai Kandiah, Dr. Matthias Haffke and Dr. Yan Nie from our laboratory for their useful input and Arturo Temblador for helping me especially with the cross-linking experiments.

The EMBL facilities were instrumental to my research, notably the Eukaryotic Expression Facility (EEF), the High-throughput Crystallization Facility (HTX) and also the PSB platforms (SEC-MALLS, Mass Spectrometry, AUC, CD etc.). Data was collected at ESRF beamlines ID29, ID23-2, and BioSAXS. Dr. Aurelien Deniaud helped with AUC experiments and Dr. Adam Round with SAXS data collection and analysis. The EMBL PhD program is acknowledged for funding.

Sincere gratitude goes to all previous and current members of the Berger and Schaffitzel laboratories, especially Alice, Arturo, Aurelien, Boris, Catarina, Christoph, Cristina, Deepak, Duygu, Etienne, Evelina, Fred, Guillaume, Isai, Jelger, Jesus, Karine, Lahari, Lakshmi, Mani, Martin, Mathieu, Matthias, Maxime, Olga, Otillie, Petra, Qiyang, Sakthi, Sarah, Simon, Taiana, Yan and all other EMBL colleagues for helpful discussions, various help and the friendly environment, and to Sarah also for French translations in this thesis.

All other friends in Grenoble and elsewhere are truly thanked for their continuous encouragement and helpful discussions.

Last but by no means least, I would like to thank my dearest family members my parents, my siblings, my wife and all other relatives for all their support throughout my PhD thesis work. Their kindness and encouragement was much appreciated, and always filled me with lots of enthusiasm to muster the effort and bring my projects to fruition.

PREFACE

Eukaryotic genomes are highly complex and can be very large. For example, the human genome contains approximately 20,000-25,000 protein coding genes. Expression of these genes needs to be tightly regulated at many levels, including chromatin organization, gene transcription, mRNA processing and export and translation, for proper functioning of cellular machinery. Many proteins and protein complexes are involved in these essential regulatory processes, examples include chromatin remodelers, transcriptional activators and coactivators, transcriptional repressors and notably the general transcription machinery. The general transcription machinery is required for and regulates the transcription of DNA into RNA. This forms the basis of the so-called central dogma of molecular biology which postulates that genetic information is realized by DNA transcription into RNA followed by translation into proteins. Transcription of protein coding genes in eukaryotes is called Class II gene transcription, and is catalyzed by RNA polymerase II (Pol II). Gene transcription by Pol II requires the cooperative interaction of multiple proteins and protein complexes to facilitate the assembly of a preinitiation complex (PIC) at the core promoter. The PIC comprises Pol II and the General Transcription Factors (GTFs)- TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, together with the Mediator complex and a large variety of transcriptional coactivators.

A fundamental step in PIC assembly is recognition of the core promoter by GTF TFIID, a megadalton sized multiprotein complex. In humans, TFIID comprises about twenty subunits made up of 14 different proteins – the TATA box binding protein (TBP) and its associated factors (TAFs, numbered 1 to 13). A range of studies on human TFIID and its subassemblies have been carried out since its discovery more than two decades ago, to understand the structure and mechanism of this essential GTF, but the architecture of TFIID, its activities, its functions, its inner workings and the mechanisms of its cellular assembly have eluded detailed understanding to date.

This thesis describes biochemical, biophysical, structural and functional studies carried out on three distinct human TFIID subassemblies. We used a number of structural biology techniques, including crystallization, nuclear magnetic resonance (NMR) spectroscopy and small angle X-ray scattering (SAXS) to analyses a complex formed by the human TBP associated factors TAF1 and TAF7. We determined crystal structures of one portion of the TAF1/TAF7 heterodimerization domain at 1.75 Å resolution, and complemented this study by determining the solution structure of a second portion of this interaction domain using NMR. We integrated these partial structures into a composite atomic model describing the entire human TAF1/TAF7 interaction by using an envelope we derived from SAXS data of the complete TAF1/TAF7 heterodimerization domain. These structural studies provide detailed insights into the intricate interaction interface formed by TAF1 and TAF7, and, together with other data available from the literature, highlight the dynamic nature of the TAF1/TAF7 interaction in the human TFIID complex.

In a second study, we analyzed a novel complex formed by TAF11, TAF13 and TBP using a range of biophysical and biochemical methods including electrophoretic mobility shift assay (EMSA), analytical ultracentrifugation (AUC), size exclusion chromatography (SEC) analysis, pull-down assay, native mass-spectroscopy and chemical cross-linking mass spectroscopy (CLMS). We discovered this complex fortuitously. Initially, based on published genetic and biochemical data, we set out to study a putative pentameric complex formed by TAF11, TAF13, General Transcription Factor TFIIA, TBP and core promoter DNA. This complex was proposed to form by interaction of the heterodimeric TAF11/TAF13 complex with a preformed TFIIA/TBP/DNA complex. In marked contrast to published data, we found that TAF11/TAF13, instead of forming a supercomplex with TFIIA/TBP/DNA, actually displaced DNA from TFIIA/TBP/DNA, and dissociated this complex. Moreover, our data evidenced that a novel, stable heterotrimeric complex composed of TAF11, TAF13 and TBP was formed, with a 1:1:1 stoichiometry in its subunits. This is reminiscent of a so-called TATA-box mimicry discovered previously in a TAF1/TBP complex. We addressed the

structure of the TAF11/TAF13/TBP complex by SAXS and CLMS, and derived models for this heterotrimeric complex based on this resolution structural information in conjunction with existing atomic structures of the subunits. These are being used to determine a quasi-atomic structural model of this interaction, as a prerequisite for *in vivo* studies to confirm the functional significance of the novel TAF11/TAF13/TBP complex we discovered, *in vitro*.

As part of the ongoing efforts in the Berger laboratory to determine the structure of human holo-TFIID, we furthermore produced and purified a large (~900 kDa) TFIID subassembly called 9TAF, which is composed of nine different TBP associated factors. 9TAF is an important subcomplex of the TFIID assembly pathway which we established *in vitro* to prepare fully recombinant 1.5 MDa human holo-TFIID complex with a full complement of TAFs and TBP. We carried out negative stain EM studies and random conical tilt (RCT) analysis on 9TAF to obtain low resolution structural information. These studies set the stage for future cryo-EM studies of this 9TAF complex in our laboratory to obtain a high(er) resolution model, for example by applying recent improvements in EM data acquisition and refinement techniques (the cryo-EM revolution), to decipher the inner workings of human TFIID.

Keywords: Transcription, pre-initiation complex PIC, General Transcription Factor GTF, human TFIID complex, TATA-box binding protein TBP, TBP associated factor TAF, Structural Biology, X-ray crystallography, small-angle X-ray scattering SAXS, nuclear magnetic resonance NMR, electron microscopy EM, electrophoretic mobility shift assay EMSA, analytical ultracentrifugation AUC, mass spectroscopy MS, chemical cross-lining mass spectroscopy CLMS.

PREFACE (EN FRANÇAIS)

Les génomes eucaryotes sont très complexes et peuvent être très grands. Par exemple, le génome humain contient environ de 20 000 à 25 000 gènes codant pour des protéines. L'expression de ces gènes doit être strictement régulée à de nombreux niveaux (tels que l'organisation de la chromatine, la transcription des gènes, le traitement et l'exportation de l'ARN messager ainsi que la traduction) pour le bon fonctionnement de la machinerie cellulaire. De nombreuses protéines et complexes protéiques sont impliqués dans ces processus essentiels de régulation, tels que les remodeleurs de la chromatine, les activateurs, co-activateurs et répresseurs de la transcription et particulièrement la machinerie générale de transcription. La machinerie générale de transcription est nécessaire et permet la régulation de la transcription de l'ADN en ARN. Cela est à la base du dogme central de la biologie moléculaire qui postule que l'information génétique est transmise par transcription de l'ADN en ARN et par traduction de l'ARN en protéines. Chez les eucaryotes, la transcription de gènes codant pour des protéines est appelée transcription génique de classe II, elle est catalysée par l'ARN polymérase II (Pol II). La transcription des gènes par la polymérase II nécessite l'interaction coopérative de plusieurs protéines et complexes protéiques afin de faciliter l'assemblage d'un complexe de pré-initiation (PIC) au promoteur de base. Le complexe de pré-initiation comprend l'ARN polymérase II et les facteurs de transcription généraux (GTFs) - TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH ainsi que le complexe de Médiateur et une grande variété de co-activateurs transcriptionnels.

Une étape fondamentale dans l'assemblage d'un complexe de pré-initiation est la reconnaissance du promoteur de base par le facteur de transcription général TFIID. TFIID est un complexe multi protéique d'environ 1,6 MDa. Chez l'homme, il comprend une vingtaine de sous-unités constituées de 14 protéines différentes - la protéine de liaison à la boite tata (TBP) et ses facteurs associés (TAFs 1 à 13). Une série d'études sur la TFIID humaine et ses sous-ensembles ont été réalisés depuis sa découverte il y a plus de 20 ans, cherchant à comprendre la structure et le mécanisme de ces facteurs de transcription général essentiel,

cependant l'architecture de TFIID, ses activités, ses fonctions, ses rouages et ses mécanismes d'assemblage cellulaire reste largement incompris à ce jour.

Cette thèse décrit les études biochimiques que nous avons effectuées sur trois sous-ensembles distincts de TFIID humain. Nous avons utilisé un certain nombre de techniques de biologie structurale : la cristallographie, la spectroscopie à résonance magnétique nucléaire (RMN) et la diffusion des rayons X aux petits angles (SAXs), pour étudier le complexe formé par les facteurs humains, associés à la protéine de liaison à la boite tata, TAF1 et TAF7. Nous avons déterminé la structures cristallisé d'une partie du domaine d'hétérodimérisation TAF1/TAF7 avec une résolution de 1,75 Å, et complété cette étude par la détermination de la structure de la solution d'une seconde partie de cette région d'interaction en utilisant la RMN. Nous avons intégré ces structures partielles dans un model atomique composite décrivant l'intégralité des régions d'interaction TAF1/TAF7 en intégrant ces résultats avec des donnés SAXS nous avons pu déterminer la structure complète du domaine d'interaction TAF1/TAF7. Ces études structurelles fournissent un aperçu détaillé sur l'interface d'interaction complexe de TAF1/TAF7, misent de concert avec des données disponibles dans la littérature, elles mettent en évidence la nature dynamique de l'interaction TAF1/TAF7 dans le complexe de TFIID humain.

Dans une deuxième étude, nous avons analysé un complexe formé par TAF11, TAF13 et TBP en utilisant un panel de méthodes biophysiques et biochimiques : l'analyse électrophorétique de retard sur gel (EMSA), l'ultracentrifugation analytique (AUC), la chromatographie d'exclusion stérique (SEC) analyse, le pull-down, la spectrométrie de masse native et la spectrométrie de masse chimique à réticulation (CLMS). Initialement, basé sur la publication de données génétiques et biochimiques, nous avons décidé d'étudier un complexe pentamérique putatif se composant de TAF11, TAF13, TBP, le facteur de transcription générale TFIIA et le cœur du complexe composé d'ADN promoteur. L'hypothèse était que ce complexe se former, par interaction de l'hétérodimère TAF11/TAF13 avec l'hétérotrimère TFIIA/TBP/ADN-promoteur. Nous avons constaté que TAF11/TAF13, au lieu de former un

supercomplexe avec TFIIA/TBP/ADN-promoteur, va en fait déplacer l'ADN de ce complexe. Nos données montrent qu'un nouveau complexe hétérotrimerique stable existe composé de TAF11, TAF13 et TBP, la stœchiométrie ses trois sous-unités étant de 1:1:1, ce trimère fait penser au complexe TAF1/TBP qui imite la boite tata. Nous avons abordé la structure du complexe TAF11/TAF13/TBP à partir de modèles dérivés de SAXS en conjonction avec des structures atomiques préexistantes des différentes sous-unités et de données de CLMS. La détermination de la structure de ce complexe est un prérequis à l'étude *in vivo* afin de confirmer le rôle fonctionnelle de ce nouveau complexe TAF11/TAF13/TBP.

De plus, dans le cadre des efforts en cours au sein du laboratoire du Pr Imre Berger afin de déterminer la structure de l'holo-TFIID humaine, nous avons reconstitué un grand sous-ensemble de TFIID (900 KDa) appelé 9TAF, qui est composé de neuf différents facteurs associés de TBP. 9TAF est un sous-complexe important de la voie d'assemblage TFIID que nous exécutons *in vitro* pour reconstituer l'intégralité du complexe holo-TFIID humain de 1,5 MDa mda. Nous avons effectué des études d'électro-microscopie par coloration négative sur le complexe 9TAF qui nous ont fourni des informations à faible résolution. Ces études ouvrent la voie à de futures études de cryo-EM sur le complexe 9TAF au sein de notre laboratoire pour obtenir un modèle de plus haute résolution, premièrement en appliquant les améliorations récentes dans l'acquisition de données EM et dans les techniques de traitement (« la révolution cryo-EM »).

Mots-cles: transcription, complexe de pré-initiation PIC, facteur de transcription général GTF, complexe TFIID humain, protéine de liaison à la boite TATA TBP, TAF facteur associé à TBP, biologie structurale, cristallographie à rayons x, diffusion des rayons x à petite-angle SAXS, résonance magnétique nucléaire RMN, microscopie électronique EM, test de décalage de mobilité électrophorétique EMSA, analytique ultracentrifugation AUC, spectroscopie de masse MS, spectroscopie de masse avec cross-linking chimiques CLMS.

ABBREVIATIONS

Å	Angstrom
A	Alanine
AA	Amino acid
Ala	Alanine
Amp	Ampicillin
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
ATP	Adenosine-5'-triphosphate
AUC	Analytical ultracentrifugation
Bdf1	Bromodomain factor 1
bp	Base pairs
β-ME	β-mercaptoethanol
BRE^d	Downstream TFIIB-recognition element
BRE^u	Upstream TFIIB-recognition element
BSA	Bovine serum albumin
BS3	Bis (sulfosuccinimidyl) suberate
C	Cysteine
Camp	Chloramphenicol
CD	Circular Dichroism
CLMS	Cross-linking mass spectrometry
Cryo-EM	Cryo electron microscopy
CV	Column volume
Cys	Cysteine
D	Aspartic acid
Da	Dalton
DCE	Downstream core element
dCTP	Deoxy cytosine triphosphate
Dmax	Maximum diameter
dNTP	Deoxynucleoside triphosphate
DNA	Deoxy ribonucleic acid
DPA	Day of proliferation arrest
DPE	Downstream promoter element
DTT	Dithiothreitol
E	Glutamic acid
EDTA	Ethylene diamine tetraacetic acid
EG	Ethylene Glycol
EM	Electron microscopy
EMDB	Electron microscopy data bank
EMSA	Electrophoratic mobility shift assay
ETO	Eight-twenty-one protein
F	Phenylalanine
FPLC	Fast protein liquid chromatography
G	Glycine
Gent	Gentamycin
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
GTF	General Transcription Factor
H	Histidine
HAT	Histone acetyl transferase
HFD	Histone fold domain
His	Histidine
hr	Hour
HTX	High-throughput crystallization
I	Isoleucine
IEX	Ion exchange chromatography
Ile	Isoleucine
IMAC	Immobilized metal affinity chromatography

Inr	Initiator
IP	Immunoprecipitation
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Initially transcribing complex
K	Lysine
Kan	Kanamycin
kDa	Kilodalton
L	Leucine
LB	Luria broth
LCR	Locus control region
Leu	Leucine
LisH	Lis homology domain
Lys	Lysine
M	Methionine
mA	milliampere
MALLS	Multi angle laser light scattering
MCS	Multi cloning site
MDa	Mega daltons
Met	Methionine
$MgCl_2$	Magnesium chloride
MHz	Mega Hertz
min	Minute
miRNA	microRNA
mM	millimolar
Mot1	Modifier of transcription1
MR	Molecular replacement
mRNA	Messenger RNA
MS	Mass spectrometry
ms	milli seconds
MTE	Motif ten element
MW	Molecular weight
N	Asparagine
Na-Acetate	Sodium acetate
Na-Citrate	Sodium citrate
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NC2	Negative cofactor 2
NMR	Nuclear magnetic resonance
nt	nucleotide
NTD	N-terminal domain
OD	Optical density
P	Proline
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PDB	Protein data bank
PEG	Polyethylene glycol
PHD	Plant homeo domain
pI	Isoelectric point
PIC	Pre-initiation complex
PolII	RNA-Polymerase II
P(r)	Pair-distance distribution
Pro	Proline
Q	Glutamine
R	Arginine
Rg	Radius of gyration
RMSD	Root mean square deviation
RNA	Ribonucleic acid
rpm	Rotation per minute
rRNA	Ribosomal RNA
S	Serine
SAGA	SPT-ADA-GCN5-acetylase
SAXS	Small angle X-ray scattering
SCP	Super core promoter
SDS	Sodium dodecyl sulfate
SEC	Size exclusion chromatography
sec	Seconds
Ser	Serine
siRNA	Small interfering RNA
SLIC	Sequence and ligation independent cloning
snRNA	Small nuclear RNA
STAGA	SPT3-TAF(II)31-GCN5L acetylase

T	Threonine
TAF	TBP associated factor
TAFH	TAF homology domain
TAND	TAF1 N-terminal domain
TATA	TATA box
TBP	TATA box binding protein
TEMED	N,N,N',N'-Tetra methyl ethylene diamine
Tet	Tetracycline
TEV	Tobacco etch virus
TFIIA	Transcription factor II A
TFIIB	Transcription factor II B
TFIID	Transcription factor II D
TFIIE	Transcription factor II E
TFIIF	Transcription factor II F
TFIIH	Transcription factor II H
TFTC	TBP-free TAF-containing complex
TLS	Translation/libration/screw
Thr	Threonine
TLF	TBP like factor
TLP	TBP like protein
TRF	TBP related factor
tRNA	Transfer RNA
Trp	Tryptophan
TSS	Transcription start site
U	Units
UV	Ultraviolet
V	Valine
V	Volt
[v/v]	Volume per volume
Val	Valine
W	Tryptophan
[w/w]	Weight per weight
[w/v]	Weight per volume
WT	Wild type
Y	Tyrosine
YFP	Yellow fluorescent protein

1. INTRODUCTION

SUMMARY IN ENGLISH

Gene expression in eukaryotes is a highly complex and tightly regulated process, which involves many protein complexes including the General Transcription Factors (GTFs), RNA polymerases, coactivators, repressors and many more, as well as defined DNA elements. Transcription of protein coding genes is called Class II gene transcription, which involves a highly regulated assembly of the pre-initiation complex (PIC) on the core promoter that is nucleated by the General Transcription Factor, TFIID.

TFIID is a large multiprotein complex of ~1.5 MDa in humans, composed of TAFs and TBP. TFIID was proposed to possess a variety of enzymatic activities functioning in reading and writing of epigenetic marks, although there is some debate in the field as to whether or not those early studies are trustworthy. TFIID was shown to recognize defined core promoter elements, such as the TATA-box, the Initiator element (Inr), the down-stream promoter element (DPE) and others. Thus, TFIID is thought to play a key role in regulating the highly complex process of gene transcription initiation. More recently, TFIID was shown to maintain the pluripotency of embryonic stem cells as well as to be also involved in maintaining and regulation of germ cells. TFIID assembly from its subunits is emerging as a key concept in regulating transcription initiation. Recent results indicated that TFIID may assemble in the cell from defined preformed subassemblies, and that nucleo-cytoplasmic transport may play a role in controlling assembly pathways. Our laboratory has determined the architecture of a ~700 kDa physiological core complex of TFIID which is present in the nuclei of cells. This nuclear TFIID core complex comprised two copies each of TAF4, TAF5, TAF6, TAF9 and TAF12. More recently, we discovered a stable subcomplex of TFIID, formed by TAF2, TAF8 and TAF10, surprisingly in the cytoplasm, which recruited importin. These findings substantiate the hypothesis of TFIID assembly from preformed stable

subunits in the cytosol and nucleus of eukaryotic cells, and indicated that indeed nuclear import may be involved in regulating this process.

To gain insight into TFIID structure and assembly, a number of structural and functional studies have been carried out over the years. EM studies of TFIID revealed the overall shape of TFIID, resembling a clamp. X-ray crystal structure and NMR solution structures of a number of isolated domains of TBP and TAFs exist. For example, many TAFs contain histone fold domains, which are protein-protein interaction motifs that also occur in nucleosomes, the fundamental unit of chromatin. TAF3, TAF4, TAF6, TAF8, TAF9, TAF10, TAF11 and TAF13 contain histone folds which prolifically occur in TFIID. These TAF histone fold domains give rise to specific TAF/TAF pairs: TAF3/TAF10, TAF8/TAF10, TAF11/TAF13, TAF6/TAF9 and TAF4/TAF12 form histone fold pairs, and atomic structures of all pairs except TAF3/TAF10 exist. Intriguingly, the structure of a dimeric complex between TBP and TAF1 has been determined by NMR and, more recently by X-ray crystallography leading to the concept of TATA-box mimicry, where a protein epitope replaces the TATA-box containing promoter DNA in the TAF/TBP complex, mimicking the DNA surface. 3D structures of isolated TAF domains exist, including the TAF1 double bromodomain and structures of the TAF5 N-terminal domain (NTD).

These studies provided a host of molecular level information about TAF interactions within TFIID, and, in the case of the TAF1 double bromodomain provided first hints about TFIID interactions with epigenetically modified histone tails in nucleosomes, illustrating the role of TFIID in transcription regulation. Progress has been markedly slow in the elucidation of TFIID structure, and for a large fraction of TFIID high resolution structural information and functional information remains elusive to date. Therefore, structural and functional studies of TFIID subcomplexes such as the studies presented in this thesis, in particular with a view to stable submodules that may represent physiological assembly intermediates, will be invaluable to unravel the structure, mechanism and genesis of TFIID in the cell.

RÉSUMÉ EN FRANÇAIS

L'expression des gènes chez les eucaryotes est un processus très complexe et étroitement régulé, qui implique de nombreux complexes protéiques tel que des facteurs de transcription généraux (GTFS), des ARN polymérases, des co-activateurs, des répresseurs et bien d'autres, ainsi que des éléments d'ADN définis. La transcription des gènes codant pour des protéines est appelée transcription des gènes de classe II, elle nécessite une fine régulation de l'assemblage du complexe de pré-initiation (PIC) sur le promoteur de base, nucléé par le facteur de transcription générale, TFIID.

TFIID est un grand complexe multiprotéique de ~ 1,6 MDa chez l'homme, composé de TAF et TBP. Des études ont proposé que TFIID ait différentes activités enzymatiques fonctionnant pour lire et écrire des marques épigénétiques, bien que la confiance en ses études soit encore discuté. TFIID reconnaît des éléments définis du promoteur de base, tels que la boîte TATA, l'élément initiateur (INR), l'élément promoteur en aval (DPE). Il a récemment été montré que TFIID permet de maintenir la pluripotence des cellules souches embryonnaires, et est également impliqué dans le maintien et la régulation des cellules germinales. TFIID est en train de devenir un concept clé dans la régulation du processus très complexe de l'initiation de la transcription des gènes. Des résultats récents ont indiqué que TFIID peut s'assembler dans la cellule à partir de sous-ensembles définis préformés, et que le transport nucléo-cytoplasmique jouerait un rôle dans le contrôle des voies d'assemblage. Notre laboratoire a déterminé l'architecture d'un complexe de base de TFIID physiologique de 700 kDa qui se trouve dans le noyaux des cellules. Ce complexe nucléaire de base de TFIID est composée de deux exemplaires de chaque TAF4, TaF5, TAF6, TAF9 et TAF12. Dernièrement, nous avons découvert un sous-complexe stable de TFIID, formé par TAF2, TAF8 et TAF10, qui sert au recrutement de l'importine ce dernier ce trouvant étonnamment dans le cytoplasme. Ces constatations viennent étayer l'hypothèse de l'assemblage de sous-unités TFIID stables préformés dans le cytosol et le noyau des cellules eucaryotes, et va dans le sens d'une régulation de ce processus par l'import nucléaire.

Pour mieux comprendre la structure et l'assemblage du complexe TFIID, un grand nombre d'études structurales et fonctionnelles ont été menées au fil des ans. Des études d'éléctro-microscopie sur TFIID ont établi la forme globale de TFIID, ressemblant à une pince. La structure par cristallographie et par RMN d'un certain nombre de domaines isolés de TBP et TAF ont été décrite. Parmi eux, de nombreux domaines TAF de repliement des histones qui sont des motifs d'interaction protéine-protéine importants pour la formation des nucléosomes, unité fondamentale de la chromatine. TAF3, TAF4, TAF6, TAF8, TAF9, TAF10, TAF11 et TAF13 contiennent des motifs structuraux de type histones nombreux dans TFIID. Ces motifs de repliement des histones des TAFs conduit à la formation de paires TAF/TAF spécifiques: TAF3/TAF10, TAF8/TAF10, TAF11/TAF13, TAF6/TAF9 et TAF4/TAF12, des structures atomiques de tous ces couples ont été publié excepté TAF3/TAF10. Etonnament, la structure d'un dimère entre TBP et TAF1 a été déterminé par RMN et, plus récemment, par cristallographie aux rayons X conduisant à la notion de « TATA-box mimétisme », où un épitope de la protéine imite la surface de l'ADN remplaçant la séquence de la TATA-box de l'ADN promoteur dans le complexe TAF/TBP. Il existe également des structure de domaines isolés TAF, parmi eux le double bromodomaine TAF1 et les structures du domaine TAF5 N-terminal (NTD).

Toutes ces études ont fourni une quantité de renseignements au niveau moléculaire sur les interactions TAF au sein de TFIID, et, le cas particulier du double bromodomaine TAF1 a procuré des premiers indices sur les interactions TFIID avec des queues d'histones portant des modifications épigénétiques dans les nucléosomes, illustrant le rôle de TFIID dans la régulation de la transcription des gènes. Les avancés ont été extrêmement lente dans l'élucidation de la structure TFIID, et encore une très grande partie de la structure de TFIID à haute résolution et de l'information fonctionnelle reste insaisissable à ce jour. Par conséquent, les études structurales et fonctionnelles de sous-complexes TFIID tels que les études présentées dans cette thèse, en particulier en vue de sous-modules stables qui peuvent

représenter des intermédiaires d'assemblage physiologiques, seront très précieux afin de démêler la structure, le mécanisme et la genèse de TFIID dans la cellule.

1.1. Overview of eukaryotic gene expression

All living organisms and also phages and viruses store genetic information in form of nucleic acid molecules — DNA and RNA. This information passes from DNA through RNA (transcription) and then from RNA to proteins (translation). This sequential flow of information is called the central dogma of molecular biology (Crick, 1958). This concept still forms the basis of much of biology, although it has been substantially expanded by the discovery of additional processes, which showed that information can also flow from RNA to DNA (Baltimore, 1970; Temin and Mizutani, 1970).

The hereditary material in eukaryotes is organized in a highly complex and compact form called chromatin, where negatively charged double-stranded DNA containing the genes is wrapped around a positively charged core octamer formed by the core histone proteins (two copies each of histone H2A, H2B, H3 and H4), to form a highly organized and dynamic protein-DNA complex called nucleosome (Kornberg, 1974; Olins and Olins, 1974; Woodcock et al., 1976). H2A-H2B and H3-H4 form heterodimer pairs via a tight structural protein-protein interaction motif which is called histone fold domain (HFD). Meanwhile, the long N-terminal extensions of the core histones (the histone tails) are unstructured and protrude out from the nucleosome (Arents et al., 1991; Luger et al., 1997). In a hallmark achievement of chromatin research, the crystal structure of the nucleosome containing 146 base-pairs of DNA wrapped around the histone octamer was determined at near-atomic resolution (Luger et al., 1997). These nucleosome particles line up together connected by linker DNA, a further histone protein called H1 and non-histone proteins, to compact in higher-order architectures summarily called chromatin, which then further condenses to make chromosomes (Horn and Peterson, 2002; Kornberg and Thomas, 1974). This structural organization of genetic material in eukaryotes is shown schematically in Figure 1.

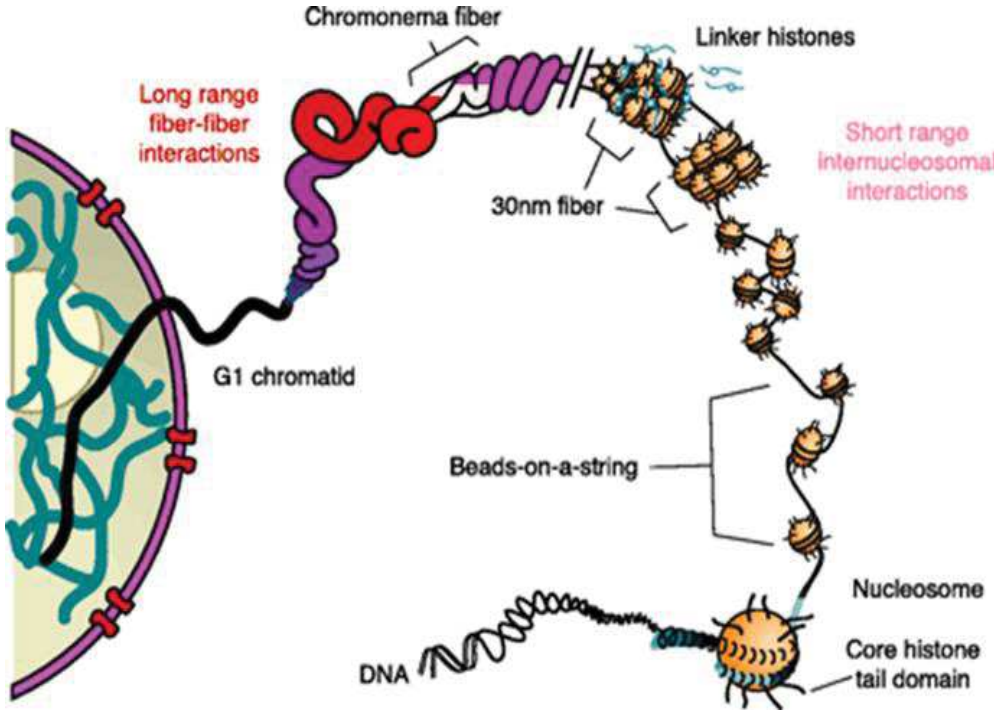


Figure 1: Structural organization of DNA to chromatin in eukaryotic nuclei.

DNA wraps around histone octamers to form nucleosomes connected by linker DNA like beads on a string, which further compacts to form the 30 nm fiber. Linker histone (H1) stabilizes these structures, which further condense to form chromonema fibers and finally chromatid. Image is from (Horn and Peterson, 2002).

Chromatin needs to be de-compacted in order to make DNA accessible to the transcription machinery to start transcription. There are two different states of chromatin found during interphase -heterochromatin and euchromatin. Euchromatin is less condensed and easily accessible "active" form, which is mainly involved in transcription of coding genes, while heterochromatin is tightly packaged and less accessible "inactive" form, which is mainly involved in chromosomal location, nuclear localization and maintaining the presence and density of repetitive DNA element (Grewal and Jia, 2007; Huisinga et al., 2006). Transition between these two states is mainly regulated by histone tail modifying enzymes and chromatin remodelers, in order to regulate transcription. Chromatin remodelers work on nucleosome substrate and catalyze various chromatin transformations, like sliding

the histone octamer across the DNA, changing the composition of the histone octamer and changing the conformation of nucleosomal DNA (Narlikar et al., 2013). Several histone tail modifications have been identified till date including phosphorylation, methylation, acetylation and ubiquitination of amino acid (AA) side chains in variable positions (Portela and Esteller, 2010) (Figure 2). Local concentration and specific combinations of these differently modified nucleosomes, so called "nucleosome or histone code", mark different epigenetic and transcriptional states (Jenuwein and Allis, 2001; Munshi et al., 2009). This complex series of process expose DNA to transcription machinery to start transcription.

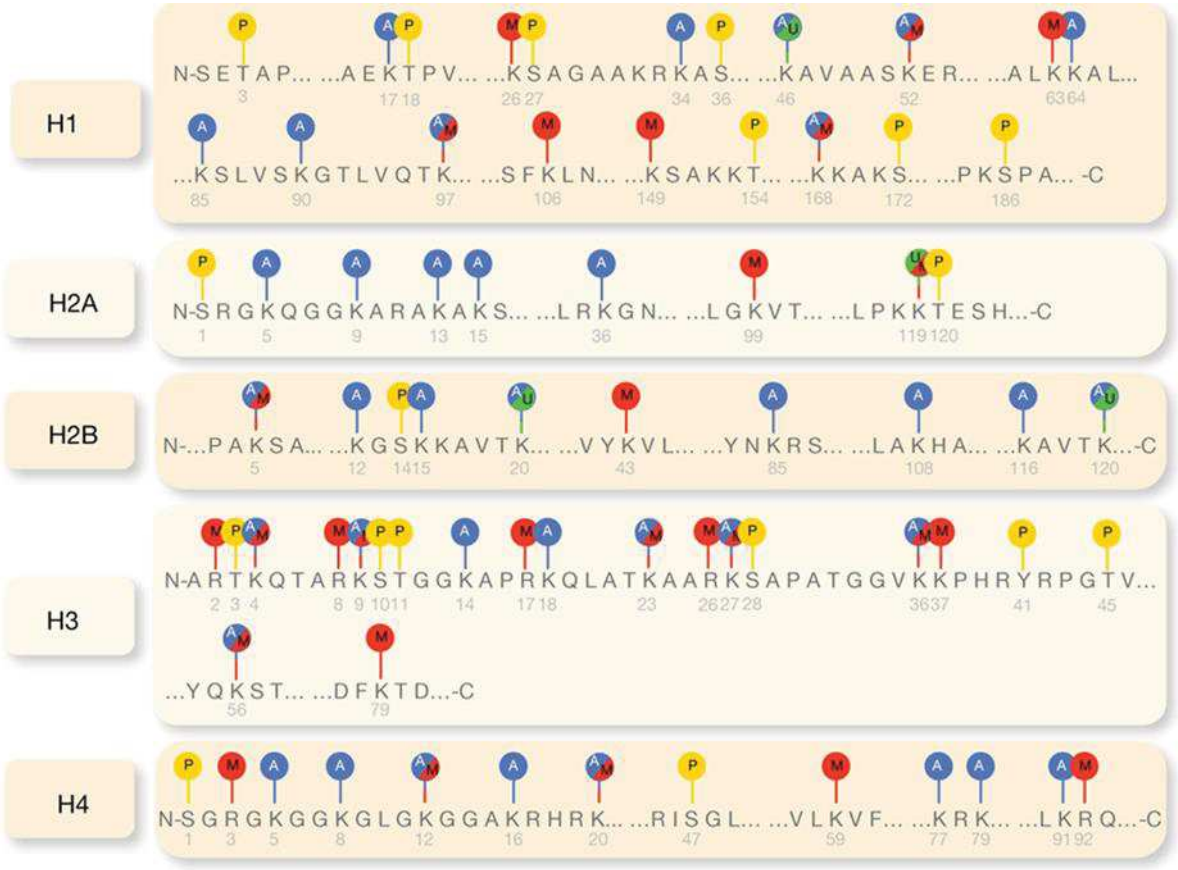


Figure 2: Histone tail modifications in eukaryotic nuclei.

Main histone modifications are shown here for histones H1, H2A, H2B, H3, and H4. Acetylation, methylation, phosphorylation and ubiquitination are shown in blue, red, yellow and green respectively. Amino acid positions are shown in gray under it. Image is from (Portela and Esteller, 2010).

Transcription is a highly coordinated process in which RNA is synthesized from DNA in a tightly regulated manner. This process involves four steps 1) Pre-initiation 2) Initiation 3) Elongation and 4) Termination. Pre-initiation and initiation are among the key regulatory steps in transcription, which occurs at specific regions on the DNA template. Once transcription has been accurately started it elongates through whole DNA template, followed by a termination signal to terminate the transcription and cleave the newly synthesized RNA (Saunders et al., 2006).

1.2. RNA Polymerases

Transcription is mediated by RNA polymerase. Enzymatic activity of the RNA polymerase was first discovered in rat liver nuclei (Weiss and Gladstone, 1959) and later found in *E. coli* as well (Chamberlin and Berg, 1962; Hurwitz et al., 1960; Stevens, 1960). Five different RNA polymerase (named I to V) have been discovered in eukaryotes, while in prokaryotes and archaea only one kind of RNA polymerase has been identified to date. Archaeal polymerase was found to be closely related (structurally and mechanistically) to eukaryotic RNA polymerase II (Korkhin et al., 2009).

In eukaryotes, RNA polymerase I is mainly involved in transcribing 18S and 28S rRNAs. RNA polymerase II is mainly catalyzes transcription of mRNA, most snRNA and miRNA. RNA polymerase III is mainly responsible for transcription of tRNAs, cellular 5S rRNA, and viral RNAs such as adenovirus VA RNAs (Kornberg, 1999; Roeder and Rutter, 1970; Sims et al., 2004; Weil and Blatti, 1976; Zylber and Penman, 1971). The recently identified RNA polymerase IV(IVa) and RNA polymerase V(IVb) are responsible for the production of siRNA in plants, mediating RNA-directed DNA methylation, transcriptional silencing, and heterochromatin formation (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005; Wierzbicki et al., 2009). All the RNA polymerases share a common function of

transcribing different DNA templates to RNA: they need assistance from other accessory protein factors to specifically recognize the transcription start sites (TSS).

1.3. Class II gene transcription

Transcription by Pol II is also known as Class II gene transcription, which is tightly regulated by the highly complex multicomponent transcription machinery. Although the transcription machinery of eukaryotes is much more complex compared to prokaryotes, the general transcription mechanism and regulation is more or less conserved. Class II gene transcription is regulated at many levels: on chromatin, during initiation and elongation of transcription, and during mRNA processing. Stability of mRNA also plays an important role in this regulation. A host of activators and repressors have been reported to regulate transcription (Venters and Pugh, 2009). A multisubunit complex, called the Mediator complex has been described as a global regulator of gene expression (Kim et al., 1994; Poss et al., 2013).

1.3.1. Core promoter elements

Class II gene transcription initiates at a defined region of DNA, the so called core promoter, which is required for proper assembly and orientation of the transcription preinitiation complex (PIC). This core promoter comprises a ~1kb long promoter region along with proximal regulatory elements, which are involved in transcription regulation (Figure 3A). A further DNA region containing distal regulatory elements is also involved in regulating transcription (Maston et al., 2006). These regulatory elements, which include enhancers, silencers, insulators or locus control regions (LCRs), are recognized by specific transcription factors to enhance or decrease transcription. Binding and functional incorporation of activators at enhancing regulatory elements has been observed to increase transcription, which results in the formation of a so-called enhancesosome (Merika et al., 1998). Similarly

a complex named repressosome made by repressors and silencing regulatory elements has been observed to decrease transcription (Gowri et al., 2003).

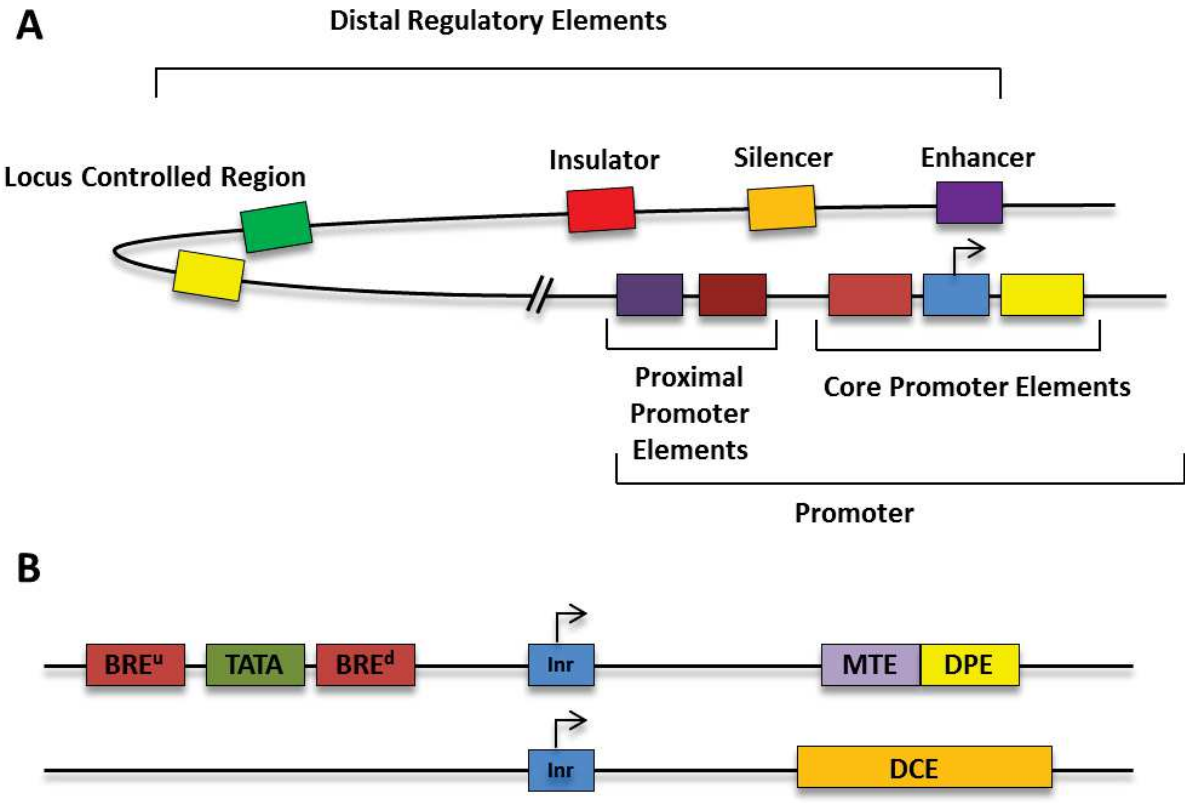


Figure 3: Regulatory DNA elements in Class II gene transcription.

(A) Different regulatory DNA elements in Class II gene transcription. The promoter (~1 kb) and different distal regulatory elements are shown here. The promoter region is composed of proximal promoter elements and the core promoter. **(B)** Different core promoter DNA elements. Upstream and downstream B-recognition elements (BRE), TATA box (TATA), initiator element (Inr), motif ten element (MTE), downstream promoter element (DPE) and downstream core element (DCE) are shown. Adapted from (Maston et al., 2006).

The core promoter contains transcription start site (+1) (TSS) and different core promoter elements- TATA box, initiator (Inr), downstream promoter element (DPE), motif ten element (MTE), downstream core element (DCE), upstream TFIIB-recognition element (BREu), downstream TFIIB-recognition element (BREd) (Figure 3B). These elements have

been shown to bind General Transcription Factors TFIID and TFIIB using electrophoretic mobility shift assays (EMSA), DNaseI foot printing and crosslinking. (Kadonaga, 2002; Kaufmann and Smale, 1994; Lee et al., 2005; Lim et al., 2004; Malik et al., 1993). The position of these from TSS, their sequences and binding proteins are shown in Table 1. TATA-binding protein (TBP) subunit of TFIID binds to TATA box (Smale and Kadonaga, 2003); TAF1 and TAF2 subunits of TFIID has been shown binding to Inr (Chalkley and Verrijzer, 1999; Verrijzer et al., 1995); TAF6 and TAF9 subunits of TFIID was shown to bind DPE (Burke and Kadonaga, 1997; Shao et al., 2005); MTE binding protein factors have not been identified yet but it works in an Inr-dependent manner, to strengthen the promoter activity (Lim et al., 2004), TAF1 subunit of the TFIID interacts with DCE (Verrijzer et al., 1995). TFIIB was shown to bind BRE^u (Lagrange et al., 1998; Littlefield et al., 1999) and BRE^d (Deng and Roberts, 2005). TAF4b/TAF12 was also shown to bind TATA box and Inr (Gazit et al., 2009). TAF4b is a paralogue of TAF4 which was isolated from B lymphocytes (Dikstein et al., 1996). These core promoter elements are present in various combinations in diverse pools of eukaryotic promoters, out of which Inr is the most common element (Gershenzon and Ioshikhes, 2005). Interestingly, comprehensive bioinformatics analysis shows that ~57% Drosophila promoter and ~68% human promoters do not contain a consensus TATA box, these are so-called “TATA-less” promoters (Kutach and Kadonaga, 2000; Suzuki et al., 2001). TATA-less promoters were already demonstrated to bind TBP, although TATA sequence was absent (Hahn et al., 1989). Based on TSS, promoters have been divided in mainly two categories: Focused promoter, where transcription starts at a single nucleotide or within a narrow region of several nucleotides; and Dispersed promoter, where transcription starts at multiple weak start sites over a broad region of about 50 to 100 nucleotides (Carninci et al., 2006; Juven-Gershon and Kadonaga, 2010). A completely different core promoter architecture has also been described, known as ATG deserts (Lee et al., 2005).

Table 1: Core promoter elements in human Class II gene transcription.

Adapted from (Thomas and Chiang, 2006).

Core Promoter Element	Position	Consensus Sequence (5' to 3')	Bound Protein in PIC
TATA	-31 to -24	TATA(A/T)A(A/T)(A/G)	TBP, TAF4/TAF12
Inr	-2 to +5	PyPyAN(T/A)PyPy	TAF2,TAF1, TAF4/TAF12
BRE^{u}	-38 to -32	(G/C)(G/C)(G/A)CGCC	TFIIB
BRE^{d}	-23 to -17	(G/A)T(T/G/A)(T/G)(G/T)(T/G)(T/G)	TFIIB
MTE	+18 to +29	C(G/C)A(A/G)C(G/C)(G/C)AACG(G/C)	not known
DPE	+28 to +34	(A/G)G(A/T)CGTG	TAF6/TAF9
DCE	3 subelements +6 to +11 +16 to +21 +30 to +34	Core sequence: S_I CTTC S_{II} CTGT S_{III} AGC	TAF1

1.3.2. General Transcription Factors (GTFs)

The requirement of accessory factors during Class II gene transcription was first shown by i*n vitro* transcription assay, where crude subcellular fractions were supplemented to accurately transcribe the native adenovirus DNA template (Weil et al., 1979). Different protein factors were found in these subcellular fractions upon further purification. These accessory factors were named TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, and were collectively defined as General Transcription Factors (GTFs). Many other activators and coactivators are also needed for accurate transcription by RNA polymerase II (Thomas and Chiang, 2006).

All these GTFs play important roles in PIC assembly and transcription initiation. Table 2 depicts the compositions and known biological function of these GTFs. Some of these are involved in recognition of specific promoter elements (see Chapter 1.3.1). TFIIA was shown to bind TBP and DNA template simultaneously to stabilize it. A structure of TFIIA/TBP/DNA shows that C-terminal domain of TFIIA binds simultaneously to TBP and

DNA to further stabilize TBP/DNA interaction (Bleichenbacher et al., 2003; Coleman et al., 1999). A TFIID subunit TAF11 was proposed to further stabilize this complex. The interaction between TFIIA and TAF11 was originally identified in a genetic screen in yeast and confirmed by biochemical means (GST-pull downs and EMSA), and TAF11/TAF13 binding was reported to stabilize the TFIIA/TBP/DNA complex (Kraemer et al., 2001; Robinson et al., 2005). TFIIB was also shown to bind to TBP and RNA polymerase II (Pardee et al., 1998; Sainsbury et al., 2013) (also see Publication 1). TFIIH and TFIIF have also been proposed to have regulatory functions (Fishburn and Hahn, 2012; He et al., 2013; Moreland et al., 1999; Murakami et al., 2013) (also see Publication 1). TFIIE is considered as the loading factor for TFIIH (Maxon et al., 1994).

Table 2: Major components of the PIC in human Class II transcription.

Adapted from (Thomas and Chiang, 2006)**.**

Factor	Protein composition	Function(s)
TFIIA	p35 (α), p19 (β), and p12 (γ)	Antirepressor; stabilizes TBP-TATA complex; coactivator
TFIIB	p33	Start site selection; stabilize TBP-TATA complex; pol II/TFIIF recruitment
TFIID	TBP + TAFs (TAF1-TAF13)	Core promoter-binding factor Coactivator Protein kinase Ubiquitin-activating/conjugating activity Histone acetyltransferase
TFIIE	p56 (α) and p34 (β)	Recruits TFIIH Facilitates formation of an initiation-competent pol II Involved in promoter clearance
TFIIF	RAP30 and RAP74	Binds pol II and facilitates pol II recruitment to the promoter Recruits TFIIE and TFIIH Functions with TFIIB and pol II in start site selection

		Facilitates pol II promoter escape Enhances the efficiency of pol II elongation
TFIIH	P89/XPB, p80/XPD, p62, p52, p44, p40/CDK7, p38/Cyclin H, p34, p32/MAT1, and p8/TFB5	ATPase activity for transcription initiation and promoter clearance Helicase activity for promoter opening Transcription-coupled nucleotide excision repair Kinase activity for phosphorylating pol II CTD E3 ubiquitin ligase activity
Pol II	RPB1-RPB12	Transcription initiation, elongation, termination Recruitment of mRNA capping enzymes Transcription-coupled recruitment of splicing and 3 end processing factors CTD phosphorylation, glycosylation, and ubiquitination

1.3.3. Preinitiation complex (PIC)

Preinitiation complex (PIC) assembly starts with binding of TFIID at core promoter followed by recruitment of GTFs and RNA polymerase II (Thomas and Chiang, 2006). A model for PIC assembly is shown in Figure 4. Initially PIC is in an inactive state and is not able to start transcription, which is called "closed" complex. Conformational rearrangements, where ~11-15 base pairs around the TSS are opened up by ATP dependent helicase activity of TFIIH, allows the single stranded DNA to position itself into the active site of RNA polymerase II, forming an "open" complex (Kim et al., 2000; Wang et al., 1992). This open complex can then further enter elongation to transcribe full length of the gene in a highly processive manner without dissociating from the DNA template or loosing the nascent RNA. Also, before fully functional elongation complex is assembled, RNA polymerase II needs to go through promoter escape or promoter clearance, in which it leaves the contacts with promoter elements and other components of the transcription machinery after synthesis of ~30 bases of RNA (Dvir et al., 1997; Hahn, 2004; Saunders et al., 2006). The components left behind and attached to promoter, remains together as a complex, which requires recruitment of TFIIB, TFIIF/RNA polymerase II only for reinitiation of transcription as proposed by (Hahn, 2004;

Yudkovsky et al., 2000). This transcription reinitiation is supposed to increase the transcription rate by escaping the slow recruitment step, in subsequent rounds of transcription. Promoter escape is preceded by an abortive transcription in many systems, where multiple short RNA products of 3-10 bases in length are synthesized (Holstege et al., 1997; Luse and Jacob, 1987).

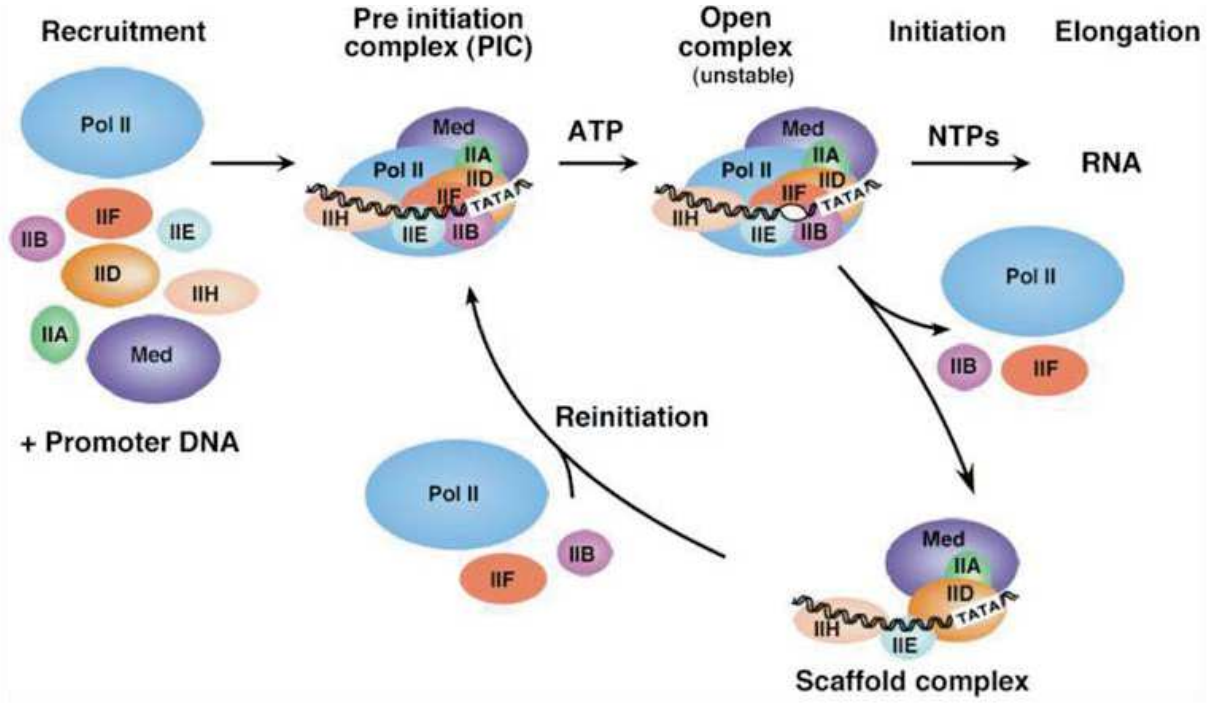


Figure 4: Model for PIC assembly and reinitiation cycle.

GTFs TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH along with Mediator (Med) and Pol II are recruited to the core promoter and PIC is formed. After start of elongation, Pol II is released from PIC and continues elongating the nascent RNA. A core scaffold complex consisting of TFIIA, TFIID, TFIIE, TFIIH and Mediator still remains bound to core promoter and help in reinitiation of transcription by recruitment of new TFIIB, TFIIF and Pol II. Image is from (Hahn, 2004).

Recent studies on human and yeast PIC formation, propose substantially different models of PIC assembly with respect to each other as wells as in comparison to current model, especially in the context of TFIIH and TFIIF. According to model for human PIC

assembly, it follows the conventional pathway except that according to this model RNA polymerase II is already recruited in the beginning without TFIIF and the function of TFIIF is to reorganize the growing PIC instead of loading RNA polymerase II and also here TFIIH is recruited in the last. On the other hand, according to the model for yeast PIC, all GTFs (except TFIIF) including TFIIH as well assemble first without RNA polymerase II, which is loaded to the assembly in last with TFIIF. These structures and models provide a significant structural and functional insight into the PIC assembly and shows that there might be different ways of PIC assembly between yeast and human (He et al., 2013; Murakami et al., 2013) (also see Publication 1). Although, recently determined architecture of yeast initially transcribing complex (ITC), an intermediate complex formed during PIC assembly by RNA polymerase II, TFIIF, TFIIB, TFIID (TBP), DNA along with a small nascent RNA (Muhlbacher et al., 2014), revealed similarities with the cryo-EM model of human ITC (He et al., 2013), which means at least core PIC is conserved between yeast and human. More recently, cryo-EM studies of ITC in complex with mediator complex provided deeper insight on transcription initiation regulation by different GTFs as well as the Mediator complex, suggesting role of mediator in stabilization of PIC as wells as in activating RNA polymerase II (Plaschka et al., 2015). Notably, in all these studies TBP was used instead of holo-TFIID, which is sufficient for only basal transcription but holo-TFIID is needed for activated transcription (Kambadur et al., 1990; Pugh and Tjian, 1990). This suggests that these PIC assembly models might be further different in activated transcription.

1.4. General Transcription Factor TFIID

TFIID (also known as TFIIDβ) is the largest and first GTF which binds to the core promoter and facilitate the recruitment of other components of PIC. It was first identified as crude chromatographic fraction (Matsui et al., 1980; Samuels et al., 1982). TFIID was shown to recognize core promoter mainly via its TATA-box binding activity and this activity was

further assigned to TFIID subunit TBP (Buratowski et al., 1988; Cavallini et al., 1988). TBP was later shown to be able to nucleate PIC (Peterson et al., 1990). Later, TFIID was shown by immunoprecipitation studies to be a multiprotein complex composed of TBP and other TBP associated factors (TAFs) (Dynlacht et al., 1991; Tanese et al., 1991). As described earlier, it was also shown that TBP alone (instead of holo-TFIID) was able to support basal transcription but failed to do so for activated transcription (see Chapter 1.3.3). Later a study showed that TBP alone was able to support some activated transcriptions like by GAL4-VP16 and GAL4-CTF1 (Oelgeschlager et al., 1998). Also, TBP has a highly conserved C terminal core domain, which was shown to be sufficient for full functionality of TBP (Cormack et al., 1991; Stevens, 1960). Different TAFs has been identified from different species including *H. sapiens, S. cerevisiae, S. pombe and D. melanogaster*, most of which are highly conserved among these species. TAFs were initially named on the basis of their particular molecular weight in individual species and later a universal nomenclature for most of the conserved TAFs was proposed, where these TAFs were named from TAF1 to TAF15, respectively from higher to lower molecular weight (Tora, 2002). Many mutational studies, which resulted in loss of active transcription has suggested the critical importance of TFIID (TBP and TAFs) (Gill et al., 1994; Pham and Sauer, 2000; Zhou et al., 1998). In humans TFIID is composed of TBP and 13 different TAFs, numbered TAF1 to TAF13. Figure 5 shows the domain representation of human TFIID components with different structural and functional domains. Histone fold domain (HFD) is a key feature of TAFs, 9 out of 13 TAFs contain these histone folds and pairs as different HFD heterodimers. All these domains are well conserved between different species. TFIID subunits were shown to be present in different copy numbers, some were present in two copies (TAF4, TAF5, TAF6, TAF9, TAF10, TAF12) and others were in single copy (TAF1, TAF2, TAF3, TAF7, TAF8, TAF11, TAF13 and TBP) (Hoffmann et al., 1996; Leurent et al., 2002; Sanders et al., 2002).

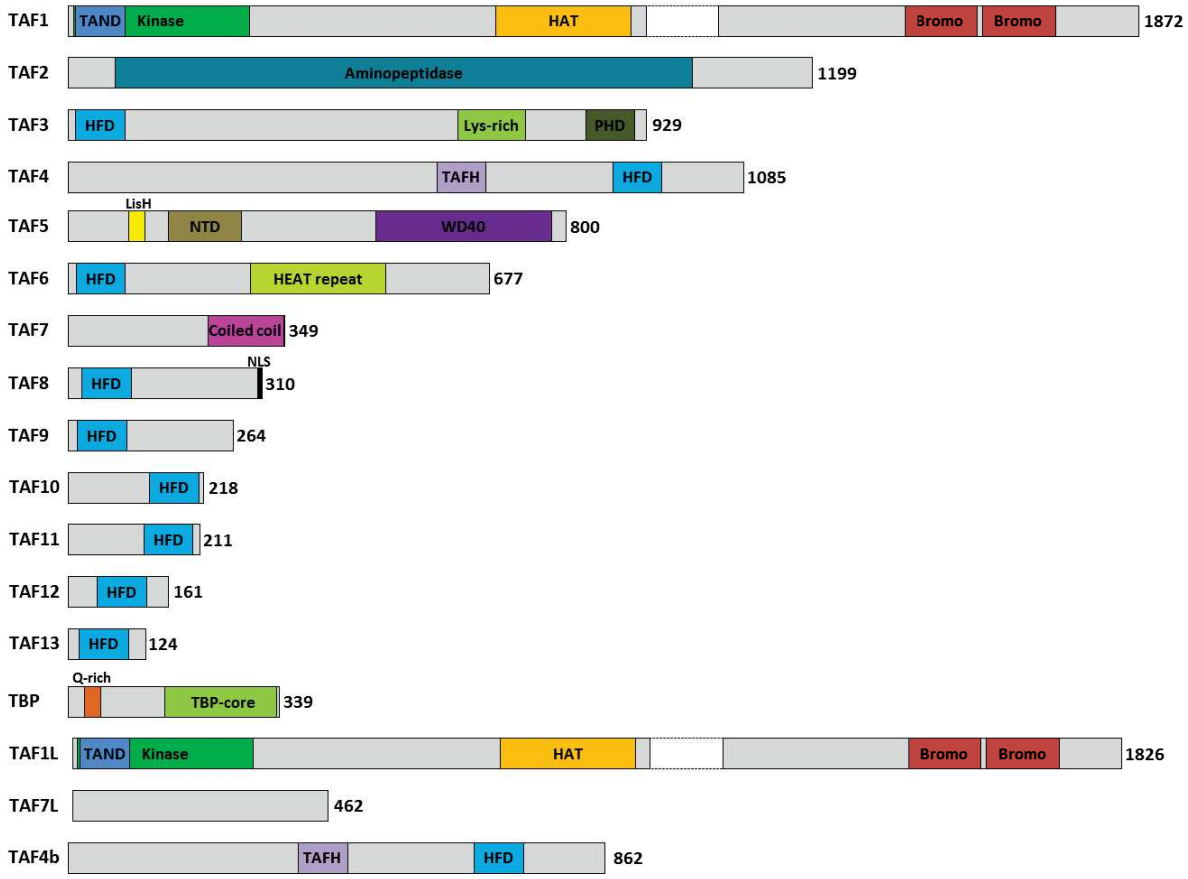


Figure 5: Subunits of human TFIID.

All subunits of the human TFIID complex are schematically represented. Paralogues of TAF1, TAF7 and TAF4 are also shown. Name of the subunit is shown on left and on right the amino acid length is shown. Different structural and functional domains and specific sequences are shown with different colors.

1.4.1. TFIID variants and isoforms

Different paralogues of TBP and TAFs have also been reported in various organisms, most of which have tissue specific functions, giving rise to different TFIID variants (Muller and Tora, 2004). For example TBP paralogues, named TBP-related proteins (TRF) or TBP-like factors (TLF) or TBP-like proteins (TLP), were discovered and was suggested to function in a similar way as TBP by binding to core promoters without consensus TATA box (Crowley et al., 1993; Hansen et al., 1997; Rabenstein et al., 1999; Wieczorek et al., 1998). Another TAFs, TAF1 and TAF7 also have male spermatogenesis specific paralogues TAF1L and TAF7L, respectively (Pointud et al., 2003; Wang and Page, 2002). TAF4b (a paralogue of

TAF4), isolated from B lymphocytes (Dikstein et al., 1996), has also been shown to be necessary for female ovarian follicle development (Falender et al., 2005b; Freiman et al., 2001) and male spermatogenesis (Falender et al., 2005a), TAF9b (a paralogue of TAF9, initially known as TAF9L) was shown to regulate neuronal gene expression other than functions similar to TAF9 (Frontini et al., 2005; Herrera et al., 2014). There are some other paralogues like TAF5b (Mitsuzawa et al., 2001), TAF10b (Georgieva et al., 2000). Two other paralogues TAF5L and TAF6L were also described (Ogryzko et al., 1998).

Most of these paralogues or isoforms were found as part of distinct TFIID complexes. Additionally, core TAFs were also shown to have varying level of expression in various tissue (Mohan et al., 2003; Perletti et al., 1999). This suggests for the presence of various TFIID variants inside the cells, which were reported to regulate and maintain local cell and tissue specific transcriptions (Muller and Tora, 2004). For example, some of the TFIID variants found in humans are, TAF10-lacking TFIID (TFIIDα) (Jacq et al., 1994), TAF4b-containing TFIID (Dikstein et al., 1996), TAF1L and TAF7L-containing TFIID (Pointud et al., 2003; Wang and Page, 2002), TAF9- lacking and TAF6δ-containing TFIID (Bell et al., 2001). Some other TFIID variants and other complexes like TAFs-lacking and BTAF1-containing TFIID (known as B-TFIID, similar to TBP/Mot1 complex, which contains yeast Mot1) (Poon et al., 1994; Timmers et al., 1992), NC2/TBP complex (Cang and Prelich, 2002; Meisterernst and Roeder, 1991), TBP-TFIIA-containing complex (TAC) (Mitsiou and Stunnenberg, 2000), TBP-free TAF-containing complex (TFTC) (Brand et al., 1999; Wieczorek et al., 1998), SPT-ADA-GCN5 acetylase (SAGA) complex, PCAF/GCN5 complex and SPT3-TAF(II)31-GCN5L acetylase (STAGA) complexes (Martinez, 2002; Muratoglu et al., 2003) were also reported to be involved in transcription in a cell/tissue/promoter specific manner.

1.4.2. Functional properties of TFIID

Different TAFs has been shown to have various structural and functional properties as shown in Table 3. As described before (see Chapter 1.3.1), TBP and some TAFs recognize and interact with core promoter elements to orient and stabilize the transcription machinery, in the beginning of transcription initiation. TAFs have various other enzymatic activities as well. TAF1 has been shown to have histone modifying activities- a histone acetyltransferase (HAT) activity to acetylate histone H3 and H4 (Mizzen et al., 1996), a kinase activity to phosphorylate histone H2B (Maile et al., 2004), a histone-specific ubiquitin-activating/conjugating activity to help in monoubiquitination of linker histone H1 (Pham and Sauer, 2000). These activities of TAF1 was shown to modulate other GTFs and cofactors as well (Imhof et al., 1997; Malik et al., 1998; O'Brien and Tjian, 1998; Solow et al., 2001). Another TFIID subunit, TAF7 was shown to inhibit HAT activity of TAF1, which resulted in loss of transcription (Gegonne et al., 2001). Here TAF7 is phosphorylated by kinase activity of TAF1, which leads to TAF7 release from TFIID and thus increase in HAT activity of TAF1 (Gegonne et al., 2006; Kloet et al., 2012) (also see Chapter 1.4.4). Some of these activities were shown for specific genes and promoters only, thus appears to be promoter or gene specific (Gegonne et al., 2001; Hilton et al., 2005). TAFs were shown to recognize epigenetic marks on histones as well - TAF1 was shown to recognize acetylated lysines on histones H3 and H4 via it's double bromodomain (Jacobson et al., 2000; Kanno et al., 2004), TAF3 was shown to recognize methylated lysines on H3 via it's PHD finger (van Ingen et al., 2008; Vermeulen et al., 2007). Interestingly, many of TAFs have been shown to form histone fold (HFD) pairs with other TAFs, and were proposed to resemble the octamer architecture of nucleosomes (Hoffmann et al., 1996; Xie et al., 1996). Recently, TFIID was also shown to have an essential role in pluripotency by inducing and maintaining the transcriptional state of these cells (Pijnappel et al., 2013). TAF2 was also proposed to contain an aminopeptidase domain based on *in silico* modeling (Malkowska et al., 2013; Papai et al., 2009).

Table 3: Components of human General Transcription Factor TFIID.

TAFs	Size (kDa)	Function(s)
TAF1	212	TATA-box mimicry, bromodomains
TAF2	136	Initiator binding
TAF3	103	histone fold, pairs with TAF10, PHD finger
TAF4	110	histone fold, pairs with TAF12,TAFH domain
TAF5	86	WD40 repeat, dimerization domain
TAF6	73	histone fold, pairs with TAF9, HEAT repeats
TAF7	40	Interacts with TAF1
TAF8	34	histone fold, pairs with TAF10
TAF9	29	histone fold, pairs with TAF6
TAF10	22	histone fold, pairs with TAF3 or TAF8
TAF11	23	histone fold, pairs with TAF13
TAF12	18	histone fold, pairs with TAF4
TAF13	14	histone fold, pairs with TAF11
TBP	38	TATA box binding protein
TAF1L	207	Paralogue of TAF1, spermatogenesis specific
TAF7L	53	Paralogue of TAF7, spermatogenesis specific
TAF4b	91	TAF4 paralogue, gonad specific, required for oocyte development

1.4.3. TFIID assembly

Not much is known to date about how TFIID assembles *in vivo*. Evidence suggests that TFIID assembly is a unique and stepwise process, not just a random accretion of different

subunits (Demeny et al., 2007). Functional partial TFIID complexes containing only subsets of TAFs have been identified *in vivo*, and recently it has been suggested that they may have unique roles in transcription regulation (Cler et al., 2009; Maston et al., 2012; Muller and Tora, 2004). A TFIID sub-complex, so-called core-TFIID, consisting of TAF4, TAF5, TAF6, TAF9, TAF12 was identified in *Drosophila* nuclei (Wright et al., 2006), and its architecture was determined recently in our laboratory by cryo-EM (Bieniossek et al., 2013) (also see Chapter 1.4.4). This complex was shown to be symmetric, and transitions to asymmetry upon accretion of TAF8/TAF10, to form a putative next intermediate in TFIID assembly, the so-called 7TAF complex (Bieniossek et al., 2013) (also see Publication 1). TAF8 was previously proposed to co-import TAF10 into the nucleus as a complex, via an importin α/β-dependent pathway (Soutoglou et al., 2005), which breaks the symmetry of core-TFIID, already present in nucleus. Recently, our laboratory showed that not only TAF10, but indeed probably in addition TAF2, is also co-imported into the nucleus with TAF8, as a preformed stable TAF2/TAF8/TAF10 complex. In this study, it was also demonstrated that TAF8 mediates TAF2 incorporation into TFIID (Trowitzsch et al., 2015), resulting in a so-called 8TAF intermediate, providing further insight into nuclear TFIID assembly. These observations suggest that preformed TFIID sub-complexes may exist in different compartments in the cell, and may play a key regulatory role in TFIID assembly. Moreover, one can speculate that these subassemblies may have own, possibly holo-TFIID independent, important transcription regulatory functions. A model for TFIID assembly based on current data is shown in Figure 6.

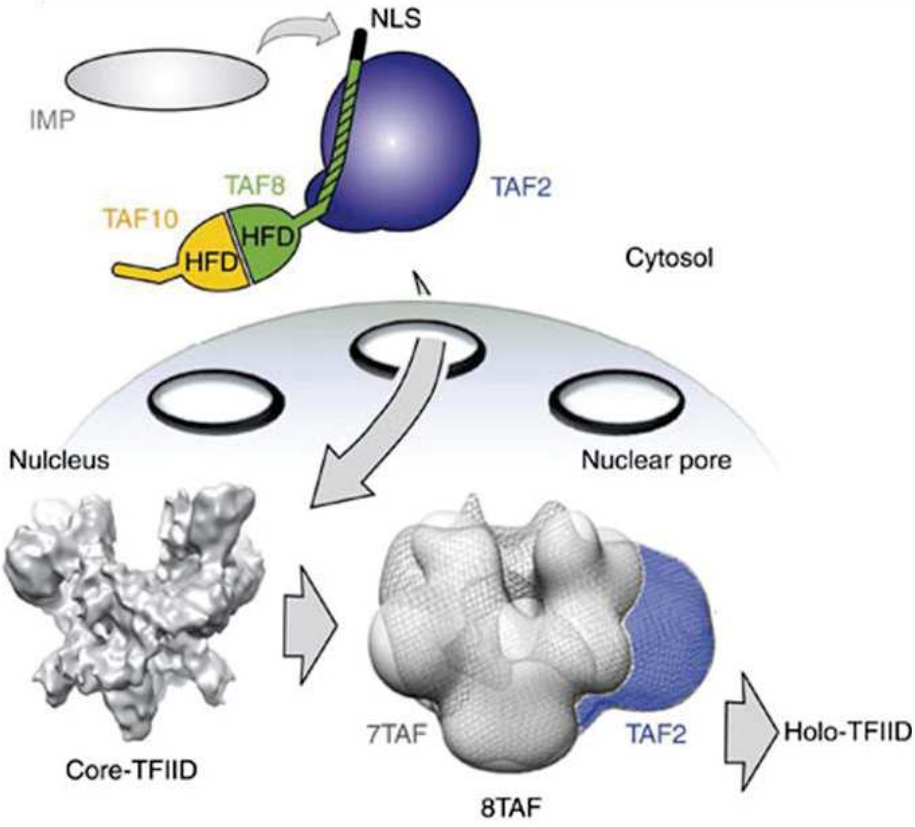


Figure 6: Model for TFIID assembly.

A TAF2/TAF8/TAF10 complex is assembled in cytosol, which is transported to nucleus via importinα/β pathway and bind to previously assembled core TFIID to form 8TAF complex, which further progresses to form holo-TFIID. Image is from (Trowitzsch et al., 2015).

1.4.4. Structural studies of TFIID

Structural characterization of various protein and protein complexes provide a host of information about biomolecular interactions (protein-protein, protein-nucleic acid), which play essential roles in biological reactions such as transcription. Acquiring this information for megadalton-sized, very scarce complexes like TFIID is rather challenging. Many structural studies have been carried out on TFIID and its components to date. Low resolution maps of TFIID were determined by negative stain-EM and cryo-EM (~22-25 Å for yeast TFIID and ~30-40 Å for human TFIID), revealing a clamp-shaped structure with three almost equal sized lobes called lobe A, B and C. Moreover, these studies also suggested a

considerable degree of conformational flexibility in TFIID, with at least two (open and closed) distinct conformations (Andel et al., 1999; Brand et al., 1999; Elmlund et al., 2009; Grob et al., 2006). Low-resolution negative stain EM envelopes of TFIID-TFIIB and TFIID-TFIIB-TFIIA also indicated the location of TFIIA and TFIIB in relation to TFIID in these supercomplexes (Andel et al., 1999). TBP and different TAFs were mapped onto TFIID using immunolabelling, followed by cryo-EM or negative stain EM (Andel et al., 1999; Leurent et al., 2002; Leurent et al., 2004; Papai et al., 2009). TBP was mapped in the central region of TFIID on lobe C, arranged between lobe A and B (Andel et al., 1999), consistent with a recent cryo-EM studies of human TFIID in presence of core promoter (Cianfrocco et al., 2013). A current cryo-EM model of TFIID is shown in Figure 7A. TAF1 was mapped at various positions in lobe A and lobe C, suggesting that it connects lobe A and lobe C spanning a rather large distance. The N-terminal domain of TAF1, known as TAND, was localized close to TBP, suggesting an interaction between TBP and TAND (Leurent et al., 2004). This TBP/TAND interaction was further explained by TBP/TAND structures from yeast and *Drosophila* (Anandapadamanaban et al., 2013; Liu et al., 1998; Mal et al., 2004) (also see Chapter 1.4.4 and Publication 1). The N-terminus of TAF2 was mapped to lobe C, whereas the C-terminus was mapped to lobes A and C close to TBP (Leurent et al., 2004; Papai et al., 2009). The NTD of TAF5 was mapped to lobe C and its WD40-repeat was located in lobes A and B (Leurent et al., 2004), confirming the presence of two copies of TAF5 as suggested previously (Sanders et al., 2002). Later, biophysical studies of isolated TAF5-NTD and TAF5-LisH/NTD constructs suggested that TAF5 may dimerize via its LisH domain (Bhattacharya et al., 2007). HFD of heterodimers TAF4/TAF12, TAF6/TAF9, TAF8/TAF10 and TAF11/TAF13 were mapped near TAF5 (Leurent et al., 2004) (Figure 7B).

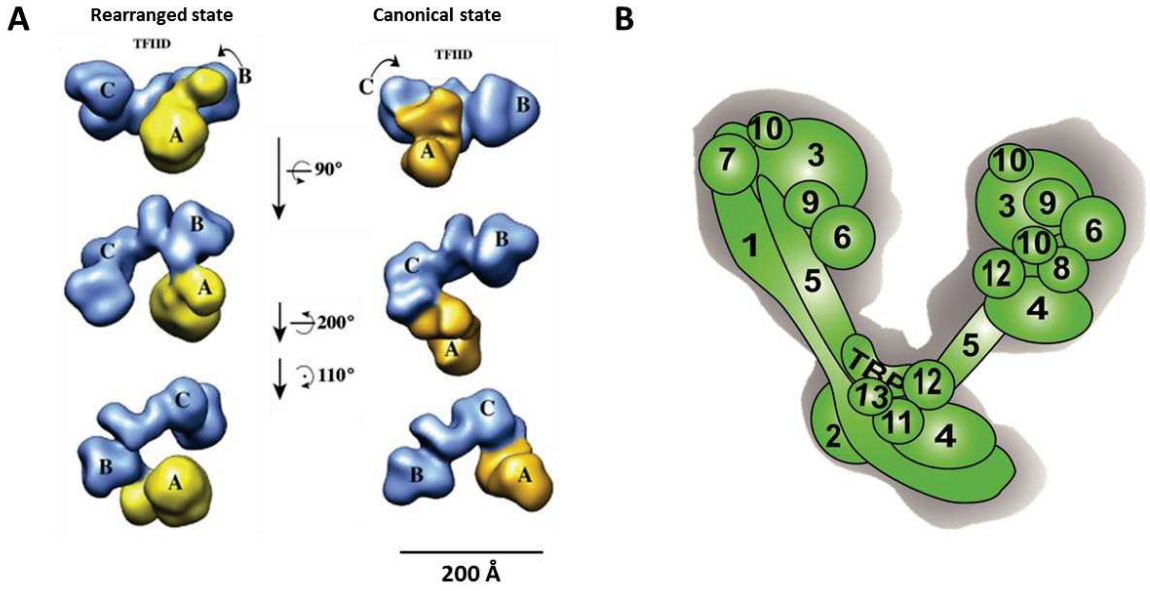


Figure 7: Overall structure of human TFIID and subunit locations.

(A) Current cryo-EM model of holo-TFIID at 35 Å resolution in canonical state as well as reorganized state, which is formed by the presence of TFIIA and promoter DNA. All three different lobes A, B and C are shown, modified from (Cianfrocco et al., 2013). **(B)** Schematic representation of location of different TFIID subunits based on different immunolabelling and EM studies.

Further cryo-EM structures of TFIID in presence of TFIIA and/or a synthetic "super core promoter" containing the DPE, MTE, Inr and TATA box, indicated that this promoter DNA may bind to a completely reorganized conformational state of TFIID (Cianfrocco et al., 2013) (also see Publication 1).

Recently, the architecture of human core-TFIID was determined by hybrid methods including cryo-EM, x-ray crystallography, homology models, mapping and proteomics, at nanometer (11.6 Å) resolution, revealing a two-fold symmetric, highly intertwined molecule with large solvent channels. This symmetric structure was shown by EM to become asymmetric upon binding of TAF8/TAF10, forming 7TAF complex. 7TAF structure was determined at ~14.3 Å resolution and a difference map of core-TFIID and 7TAF showed rearrangement of different lobes and an extra density, which was attributed to TAF8/TAF10

binding by EM analysis of 7TAF using a TAF10 specific antibody. Known structures of conserved domains of the TAFs, and homology models, were placed in the core-TFIID cryo-EM envelope (Bieniossek et al., 2013) (Figure 8A). This study was consistent with findings of the previous studies, which had indicated that two copies each of TAF5-NTD and TAF5-WD40 repeat domains were distantly positioned from each other and, moreover, the HFD pairs of TAF6/9 and TAF4/12 were placed close to TAF5.

An 8TAF complex (an intermediate in TFIID assembly formed by adding TAF2 to the 7TAF complex) was also analyzed by negative stain EM, revealing the location of TAF2 in 8TAF (Trowitzsch et al., 2015) (Figure 8B). All these studies in their aggregate provide structural information of the interactions within core-TFIID and provide some clues of the further assembly into holo-TFIID. Furthermore, the studies suggest that profound structural rearrangements are occurring during the accretion of TAFs and TBP, possibly leading to the provision of subsequent binding sites for further TAFs in the holo-TFIID assembly pathway.

Recently, the architecture of yeast SAGA complex, a large transcriptional activator complex which shares subunits with TFIID, was modelled using a cross-linking mass spectrometry based approach, revealing various interactions between its components and confirming the structural arrangements of the shared TAFs that were previously found in core-TFIID, indicating architectural similarities (Han et al., 2014).

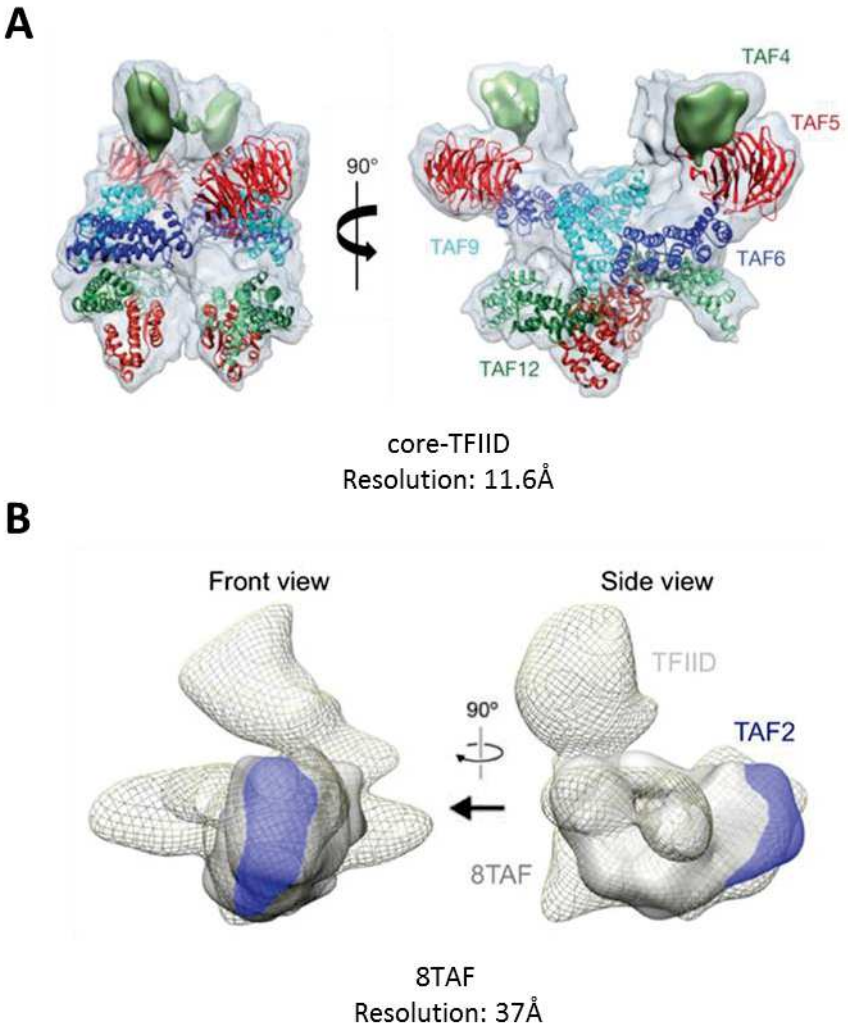


Figure 8: Cryo-EM and negative stain EM models of TFIID sub-complexes.

(A) Cryo-EM model of core-TFIID complex is shown from different orientations. Fitting of high resolution crystal structures (represented as ribbons), within these models is also shown. TAF5, TAF6 and TAF12 domains are shown in red, blue/cyan and in green respectively. Modified from (Bieniossek et al., 2013). **(B)** Negative stain EM model of 8TAF (grey) at low (37 Å) resolution, superimposed on a cryo-EM model of holo-TFIID (EMD-1195) (grey mesh) is shown, with TAF2 position colored in blue. Adapted from (Trowitzsch et al., 2015).

Other than holo-TFIID and its assembly intermediates, atomic resolution structures of conserved domain(s) of TBP as well as of component TAFs alone or in complex with other TAFs, TFIIA, TFIIB and/or TATA-box DNA, especially HFD pairs, were also determined,

using mainly X-ray crystallography and NMR. Atomic resolution X-ray structures of the conserved TBP core were determined in isolation and in complex with many TATA-box DNAs from various organisms, revealing a convex surface and a concave surface where DNA binds. Interestingly free TBP forms a tight dimer (Chasman et al., 1993; Kim et al., 1993b; Nikolov et al., 1996) (Figure 9C). Molecular structures of TBP in complex with DNA and TFIIA were also determined to identify the atomic details of how this GTF enhances the stability of the TBP/DNA complex (Bleichenbacher et al., 2003; Tan et al., 1996) (Figure 9C). Structures of TFIIB/TBP/DNA complexes were also determined (Nikolov et al., 1995). Structures of TBP with negative cofactors NC2 along with DNA explained how this cofactor regulates transcription (Kamada et al., 2001). Further, the crystal structure of TBP in complex with mot1 (yeast homologue of human BTAF1) (Wollmann et al., 2011) and crystal and NMR structures of TBP/TAF1-TAND (TAF N-terminal Domain) from yeast and *Drosophila* (Anandapadamanaban et al., 2013; Liu et al., 1998; Mal et al., 2004), explained a new phenomenon in transcriptional regulation called TATA-box mimicry, in which these domains bind to the DNA-binding concave surface of TBP and regulate binding of DNA and different negative and positive transcription factors to TBP in order to regulate transcription (Figure 9C).

TFIID is abundant in HFD pairs, which were proposed to form histone like octamers or tetramers within TFIID, based on structural and biochemical studies of the TAF4/TAF12 and TAF6/TAF9 heterodimers from yeast and *Drosophila*. Promoter DNA was proposed to wrap around these octamers in TFIID, just like in the nucleosomes (Hoffmann et al., 1996; Oelgeschlager et al., 1996; Selleck et al., 2001; Xie et al., 1996). However, the HFDs of TFIID don't contain the basic amino acid side chains as in the histones, which are required for stable interaction with the DNA. Structures of HFDs of TAF6/TAF9 from drosophila and TAF11/TAF13, TAF4/TAF12 and TAF8/TAF10 from human have been determined, which were structurally similar to HFDs of nucleosome histone heterodimers H3-H4 and H2A-H2B (Birck et al., 1998; Trowitzsch et al., 2015; Werten et al., 2002; Xie et al., 1996) (Figure 9A).

Structural information on only TAF3/TAF10 HFD is still unavailable. There is a possibility that the interaction surface of TAF3 and TAF8 with TAF10 might have specific differences, which may explain any functional significance of pairing of TAF10 with two different TAFs.

Different other conserved domains of TAFs have also been studied at atomic resolution as shown in Figure 9B. Structural characterization of double bromodomain (DBD) from human TAF1 along with biochemical studies showing a strong binding of the bromodomain module to mono- and di-acetylated but not to unmodified histone H4 tails makes TAF1 a reader of epigenetic marks (Jacobson et al., 2000). Interestingly, yeast double bromodomain is present in a so called Bromodomain factor 1 (Bdf1) instead of TAF1, which binds tightly with TFIID and is thought to compensate for the missing C-terminal domain of yeast TAF1 (Matangkasombut et al., 2000). Bromodomains of human TAF1 was also determined along with of the paralogue human TAF1L for large scale studies of bromodomain family, which revealed conservation of bromodomain in TAF1 and TAF1L (Filippakopoulos et al., 2012). Further, the structure of TAF5 NTD from yeast, human and *E.Cuniculi* were also determined, which was mostly formed of α-helices but with less similarities to other α-helical bundles containing structures (Bhattacharya et al., 2007; Romier et al., 2007). Structural characterization of human TAF4-TAFH domain showed structural similarities with the eight-twenty-one (ETO) protein ETO-TAFH domain. This TAF4-TAFH domain was shown to preferably bind to a short hydrophobic peptide motif by binding cleft analysis and peptide phage display. This hydrophobic peptide is also found in a number of transcriptional regulators, because of which this domain was proposed to have a specific co-activator function within TFIID (Wang et al., 2007; Wei et al., 2007). A highly conserved HEAT repeat domain of TAF6 was also explained structurally from *A. Locust* (Scheer et al., 2012) and Human (Matthias Haffke, Berger group, unpublished). NMR solution structures of mouse TAF3-PHD finger domain alone and in complex with H3K4me3 peptide along with mutational and biophysical studies revealed the role of this domain in recognition of epigenetic marks (van Ingen et al., 2008).

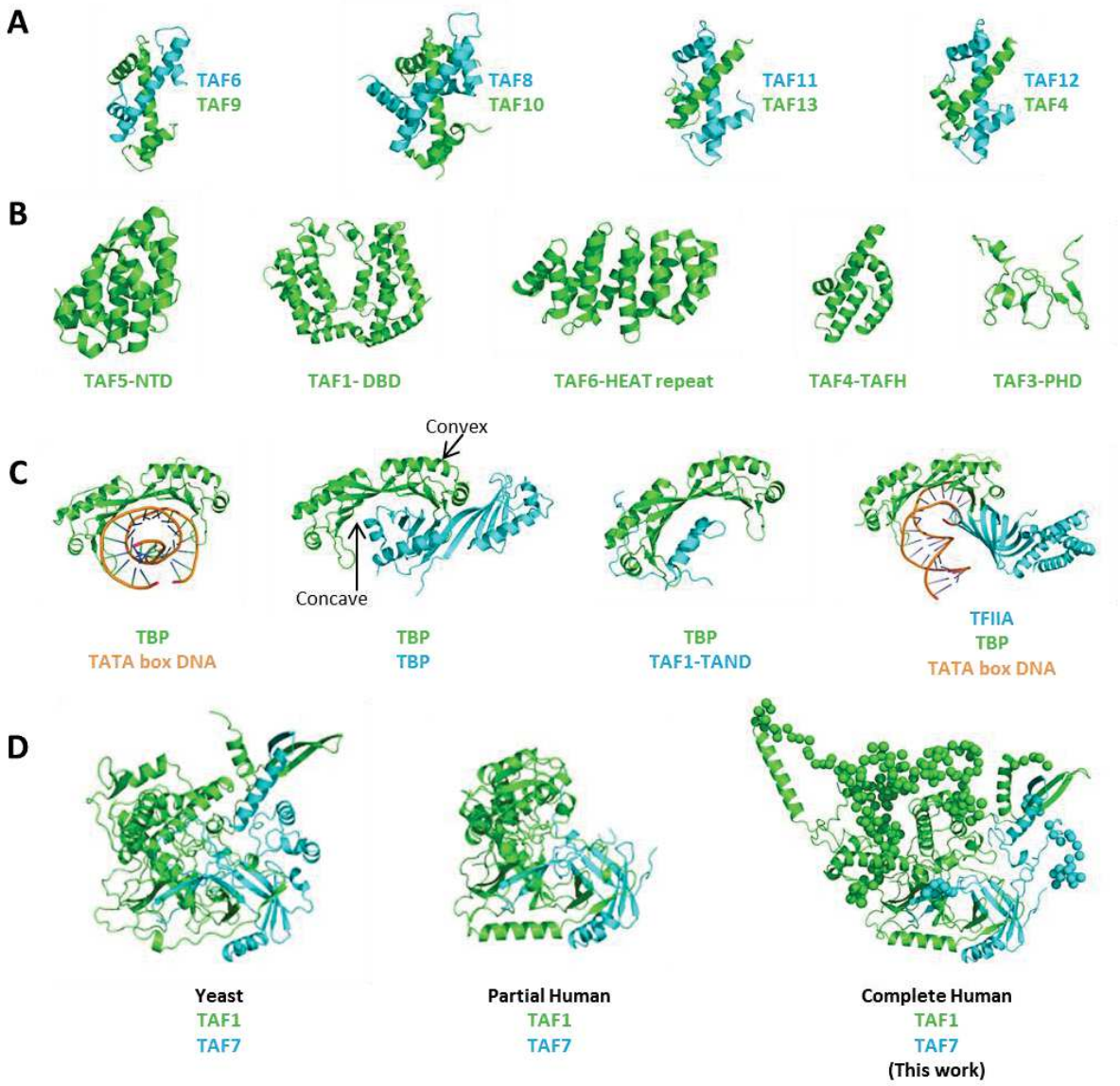


Figure 9: Crystal and NMR structures from TFIID subunits.

(A) Crystal structures of the HFD pairs: drosophila TAF6/TAF9 (PDB ID 1TAF) (Xie et al., 1996), human TAF8/TAF10 (PDB ID 4WV4) (Trowitzsch et al., 2015), human TAF11/TAF13 (PDB ID 1BH8) (Birck et al., 1998) and human TAF4/TAF12 (PDB ID 1H3O) (Werten et al., 2002). **(B)** Crystal structures of individual domains of different TAFs: human TAF5-NTD (PDB ID 2NXP) (Bhattacharya et al., 2007), human TAF1-DBD (double bromodomain) (PDB ID 1EQF) (Jacobson et al., 2000), *A. locustae* TAF6-HEAT repeat (PDB ID 4ATG) (Scheer et al., 2012), human TAF4-TAFH (PDB ID 2P6V) (Wang et al., 2007) and NMR structure of mouse TAF3-PHD (PDB ID 2K16) (van Ingen et al., 2008). **(C)** Crystal structures of conserved core of TBP: human TBP in complex with TATA-box DNA (PDB ID 1CDW) (Nikolov et al., 1996), Yeast TBP dimer (PDB ID 1TBP) (Chasman et al., 1993), yeast TBP/TAF1-TAND

complex (PDB ID 4B0A) (Anandapadamanaban et al., 2013), human TBP/TFIIA/DNA complex (PDB ID 1NVP) (Bleichenbacher et al., 2003). Convex and concave surfaces of TBP are also shown with arrows. **(D)** Crystal structure of yeast TAF1/TAF7 complex (left) (PDB ID 4OY2) (Bhattacharya et al., 2014), incomplete human TAF1/TAF7 complex (middle) (PDB ID 4RGW) (Wang et al., 2014), and complete human TAF1/TAF7 complex (right) by hybrid methods (this work).

During the studies on the human TAF1/TAF7 interaction by hybrid methods carried out in this thesis, a crystal structure of a TAF1/TAF7 complex from yeast (Bhattacharya et al., 2014) was determined, followed shortly by a partial structure of human TAF1/TAF7 complex (Wang et al., 2014) (Figure 9D). These structures include previously proposed, putative HAT domain of TAF1, and, remarkably, no structural similarities with known HAT domains from other proteins were observed, casting some doubts on the previously assigned HAT activity of TAF1. Note that this crystal structure of the human TAF1/TAF7 interaction was incomplete as it was missing part of the interacting domain.

In summary, TFIID and its subunits (TAFs and TBP) have been intensely researched since the discovery of this essential GTF. We have now a reasonable view on the overall shape of TFIID at low resolution, derived from EM studies. We also have gained first substantial insight into TFIID assembly by the discovery of stable preformed submodules in the nucleus (core-TFIID) and the cytosol (TAF2/TAF8/TAF10 complex). High resolution structure of parts of the TFIID subunits have been determined, although progress has been slow, and the studies have been predominantly limited to the histone fold pairs within TFIID which however mostly only constitute small proportions of the TAFs that contain the histone fold motifs. Insight about the supramolecular arrangement of TAFs and TBP in TFIID, and the rearrangements during TFIID assembly, are emerging. Nonetheless, intense efforts and investments are required to progress towards a complete atomic understanding of TFIID architecture, its dynamic assembly (notably also with a view to paralogues and isoforms including tissue specific forms of TFIID) and their interactions with chromatin and the other components of the Class II transcription machinery.

1.5. Aims of this thesis

TFIID has a multitude of roles in transcription regulation, in maintaining cell differentiation, development and homeostasis in eukaryotes. Many studies have been carried out; however, a very substantial body of structural and functional information remains ahead of us to achieve a complete atomic understanding of TFIID structure, function and cellular assembly. In the Berger laboratory, we have achieved the reconstitution of fully recombinant human holo-TFIID with a full complement of TAFs and TBP, which is active in transcription assays and competent in chromatin binding PhD thesis of Yan Nie (Nie, 2012). Our current protocol for holo-TFIID production includes a reconstitution step *in vitro*, involving reaction of a so-called 9TAF complex (comprising TAF2, TAF3, TAF4, TAF5, TAF6, TAF8, TAF9, TAF10, TAF12) to a module which we termed the MBPTAF1-Module (comprising TAF1, TAF7, TAF11, TAF13 and TBP). Structural and functional studies of these separate components are major objectives of this thesis to obtain insights into the inner workings of holo-TFIID, with the overarching aim to further our understanding of TFIID structure and function and the mechanisms of cellular TFIID assembly.

In this thesis, I will describe studies of three different subassemblies of human TFIID. In the first part of this thesis, I carried out structural and dynamic studies of a human TAF1/TAF7 complex containing the complete TAF1/TAF7 dimerization interface by hybrid methods. In the second part of this thesis, I describe structural and functional studies of a novel TAF11/TAF13/TBP complex which we discovered. Part I and Part II of my thesis address components of our so-called MBPTAF1-Module. Part III of this thesis finally described structural studies by cryo-EM of recombinant 9TAF complex, which is the complementary constituent of our *in vitro* TFIID assembly method.

PUBLICATION 1

More pieces to the puzzle: Recent structural insights into Class II transcription initiation.

Authors: *Eaazhisai Kandiah, Simon Trowitzsch,* ***Kapil Gupta****, Matthias Haffke and Imre Berger*

Current Opinion in Structural Biology; Volume 24, February 2014, Pages 91–97
DOI: 10.1016/j.sbi.2013.12.0

This paper reviews recent structural studies of many of very large assemblies of the PIC, containing more than 30 proteins and promoter DNA, to understand the Class II gene transcription initiation in more details. Furthermore, this review explains how improvements in recombinant expression techniques, *in vitro* reconstitution approaches and powerful integrated structure determination methods have helped in obtaining this structural information.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

More pieces to the puzzle: recent structural insights into class II transcription initiation

Eaazhisai Kandiah[1,2], Simon Trowitzsch[1,2], Kapil Gupta[1,2], Matthias Haffke[1,2] and Imre Berger[1,2]

Class II transcription initiation is a highly regulated process and requires the assembly of a pre-initiation complex (PIC) containing DNA template, RNA polymerase II (RNAPII), general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH and Mediator. RNAPII, TFIID, TFIIH and Mediator are large multiprotein complexes, each containing 10 and more subunits. Altogether, the PIC is made up of about 60 polypeptides with a combined molecular weight of close to 4 MDa. Recent structural studies of key PIC components have significantly advanced our understanding of transcription initiation. TFIID was shown to bind promoter DNA in a reorganized state. The architecture of a core–TFIID complex was elucidated. Crystal structures of the TATA-binding protein (TBP) bound to TBP-associated factor 1 (TAF1), RNAPII–TFIIB complexes and the Mediator head module were solved. The overall architectures of large PIC assemblies from human and yeast have been determined by electron microscopy (EM). Here we review these latest structural insights into the architecture and assembly of the PIC, which reveal exciting new mechanistic details of transcription initiation.

Addresses
[1] European Molecular Biology Laboratory, Grenoble Outstation, 6 rue Jules Horowitz, 38042 Grenoble, France
[2] Unit of Virus Host-Cell Interactions, Unité Mixte Internationale UMI 3265, Université de Grenoble Alpes – EMBL – CNRS, 6 Rue Jules Horowitz, BP181, 38042 Grenoble Cedex 9, France

Corresponding author: Berger, Imre (iberger@embl.fr)

Current Opinion in Structural Biology 2014, **24**:91–97

This review comes from a themed issue on **Nucleic acids and their protein complexes**

Edited by **Karolin Luger** and **Simon EV Phillips**

For a complete overview see the Issue and the Editorial

Available online 16th January 2014

0959-440X/$ – see front matter,

http://dx.doi.org/10.1016/j.sbi.2013.12.005

Introduction

Transcription of RNAPII dependent genes involves a cascade of events including binding of activators to enhancers, assembly of the PIC on promoters and elongation by RNAPII [1]. PIC assembly is synergistic and highly regulated, involving large multiprotein complexes such as RNAPII, TFIID, Mediator and others [2]. TFIID is thought to nucleate the PIC on a core promoter by binding to the TATA box *via* its TBP subunit [3]. Binding of TFIIA stabilizes TFIID–core promoter interactions [4,5]. TFIIB interacts with TBP and promoter DNA, and recruits the RNAPII–TFIIF complex [6,7]. With TFIIB and TFIIF, RNAPII orients the DNA template and selects the transcription start site (TSS) [8]. TFIIE and TFIIH are then accreted and together with RNAPII catalyze promoter melting and transition from transcription initiation to elongation [3].

Understanding the mechanism of transcription initiation requires structural knowledge of the components and their interplay. During the last three years, a number of revealing structures of important transcription initiation subassemblies have been obtained. Here we review new studies on TFIID and its components [9•,10•,11•], the RNAPII–TFIIB initially transcribing complex (ITC) [12••] and large PIC assemblies containing RNAPII and GTFs [13••,14••]. These exciting results considerably advance our understanding of TFIID assembly and its regulatory role in promoter binding, the function of TFIIB in RNAPII transcription initiation, and the supramolecular organization of the entire PIC.

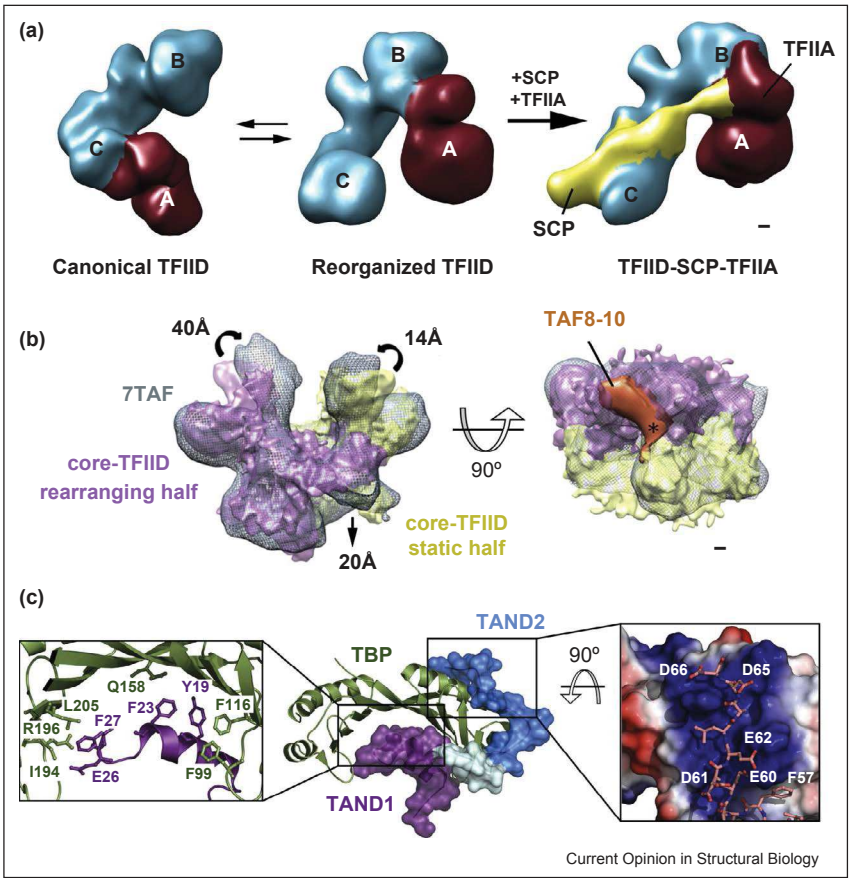
New insights into TFIID structure and assembly

TFIID is a megadalton-sized multi-subunit complex containing TBP and 13–14 TAFs [3]. Recent structural analyses by cryo-EM reveal holo-TFIID binding to promoter DNA in a reorganized state, and provide a view of a physiological core–TFIID complex with implications for holo-TFIID assembly [9•,10•].

The structure of holo-TFIID is dynamic

TFIID adopts a horseshoe-shaped structure comprising three lobes (A, B and C) (Figure 1) [5]. In its 'canonical' form, lobe A is anchored to lobe C (Figure 1a). Cryo-EM analysis of human holo-TFIID, purified from HeLa cells, revealed a hitherto unobserved 'reorganized' form [9•]. In the reorganized form, lobe A dislocates from lobe C to lobe B (Figure 1a). When bound to TFIIA and an artificial core promoter DNA (super core promoter, SCP), the reorganized form of TFIID is mainly present, in contrast to unbound TFIID which exists in both forms. SCP contains a TATA-box, initiator, downstream promoter element and motif ten element [15], and interactions of these elements with TFIID subunits could be tentatively assigned to the lobes in the TFIID–SCP–TFIIA structure [9•]. Based on molecular modeling and docking analysis the authors propose that the N-terminus of TAF1

Figure 1



TFIID architecture. **(a)** Cryo-EM studies of endogenous TFIID purified from HeLa cells reveal two distinct forms. In the canonical form (left), lobe A (colored in red) is attached to a scaffold formed by lobes B and C (colored in blue) by contacting lobe C. In the reorganized form (middle), lobe A is contacting lobe B. When promoter DNA (colored in yellow) and TFIIA are added, TFIID is stabilized in the reorganized form (right). SCP stands for super core promoter DNA [15]. The position of TFIIA (marked) was deduced from nanogold labeling experiments [9•]. Scale bar, 1 nm. **(b)** Cryo-EM structure of a human core–TFIID complex formed by two copies each of five TAFs (TAF4, 5, 6, 9 and 12) is shown in two colors (yellow and purple). The cryo-EM structure (gray mesh) of a complex (7TAF) containing two further TAFs (TAF8 and TAF10) is superimposed. Core–TFIID is symmetric. Addition of TAF8 and TAF10 results in a rearrangement of the subunits, giving rise to an asymmetric structure. The structural rearrangements (indicated by arrows) occur mainly in one half of the structure corresponding to the 'core–TFIID rearranging half' (colored in purple), while the 'core–TFIID static half' (colored in yellow) remains largely unaltered (left panel). A bottom view (right panel) shows the density corresponding to the TAF8–10 heterodimer (colored in orange). The location of the twofold axis relating the two halves in core–TFIID is marked by an asterisk [10•]. Scale bar, 1 nm. **(c)** Crystal structure of the TAND1 (colored in purple) and TAND2 (colored in blue) domains of yeast TAF1 bound to TBP (colored in green). The TAND1 and TAND2 domains are depicted in a surface representation; TBP is shown as a ribbon (middle panel). TAND1 and TAND2 are connected by a linker (colored in light blue). Mainly hydrophobic interactions are found at the interface between TAND1 and TBP (left panel). Mainly ionic interactions are found at the interface between TAND2 (residues shown as sticks) and TBP (in electrostatic surface representation) (right panel). A phenylalanine (F57) serves as an aromatic anchor [11•].

is present in lobe A, whereas the C-terminal portion of TAF1 reaches over onto the preassembled BC lobes, suggesting a structural role for TAF1 in the transition between the canonical and reorganized forms of TFIID [9•,16].

Architecture of a physiological core–TFIID

A subset of TAFs (TAF4, 5, 6, 9, 10 and 12) within TFIID are thought to be present in two copies, while the remaining TAFs (1, 2, 3, 7, 8, 11 and 13) and TBP are present in single copy [17]. This arrangement conveys a bipartite architecture for TFIID and a transition from a symmetric to an asymmetric structure during the assembly of TFIID. A physiological core complex of TFIID was identified in *Drosophila* nuclei, containing a subset of five TAFs, TAF4, 5, 6, 9 and 12 [18]. This core–TFIID complex was proposed to function as a central scaffold for the binding of the remaining TAFs and TBP on the periphery [18].

The architecture of fully recombinant core–TFIID from human, composed of TAF4, 5, 6, 9 and 12, was recently

analyzed using a hybrid approach involving cryo-EM, data from X-ray structures and homology modeling [10•]. This study revealed that these TAFs, present in two copies each, assemble into a twofold symmetric structure (Figure 1b). Conserved domains within the TAFs forming this core–TFIID, including the histone-fold (HF) domains of TAF 4, 6, 9 and 12 were placed in the EM density [10•]. The HF is a prevalent structural feature within TFIID, with 9 of 14 TAFs containing a HF domain [19]. The presence of HF domains led to the proposal that TFIID may contain histone tetramer and octamer-like structures, as they occur in nucleosomes. The core–TFIID cryo-EM structure, comprising the TAF6–9 and TAF4–12 HF pairs, however, did not provide evidence for this hypothesis [10•]. Whether subsequent conformational changes induced by the addition of further TAFs containing HF domains may result in octamer-like structure remains to be elucidated.

Extensive conformational changes were observed, when a complex of core–TFIID bound to two HF domain containing TAFs, TAF8 and TAF10, was analyzed by cryo-EM (Figure 1b). This 12 subunit complex contains two copies each of TAF4, 5, 6, 9 and 12, but only one copy each of TAF8 and TAF10, and, in contrast to core–TFIID, is asymmetric [10•]. The binding of one TAF8–10 heterodimer appears to rule out the binding of a second, by means of steric hindrance (Figure 1b). Proteomics studies confirmed that the stoichiometry found in the recombinant TAF complexes is consistent with those present in endogenous TFIID [10•]. Taken together, these results provide a mechanistic model for the structural transition of TFIID from a symmetric core formed by TAF4, 5, 6, 9 and 12, to an asymmetric holo-TFIID complex, triggered by binding of TAF8–10 in single copy and involving major structural rearrangements in the core–TFIID scaffold [10•].

Crystal structure of a yeast TAF1–TBP complex

TAF1 interacts strongly with TBP, and mutations modulating TAF1–TBP binding affect cell growth and division [20]. TAF1–TBP interactions were mapped to the TAF N-terminal domain (TAND) region of TAF1 [20]. A structural model for the TAF1–TBP interaction was deduced from earlier nuclear magnetic resonance (NMR) data, suggesting that the N-terminal region of TAF1 regulates TBP activity by mimicking TATA-box DNA and occupying the DNA binding surface of TBP [21].

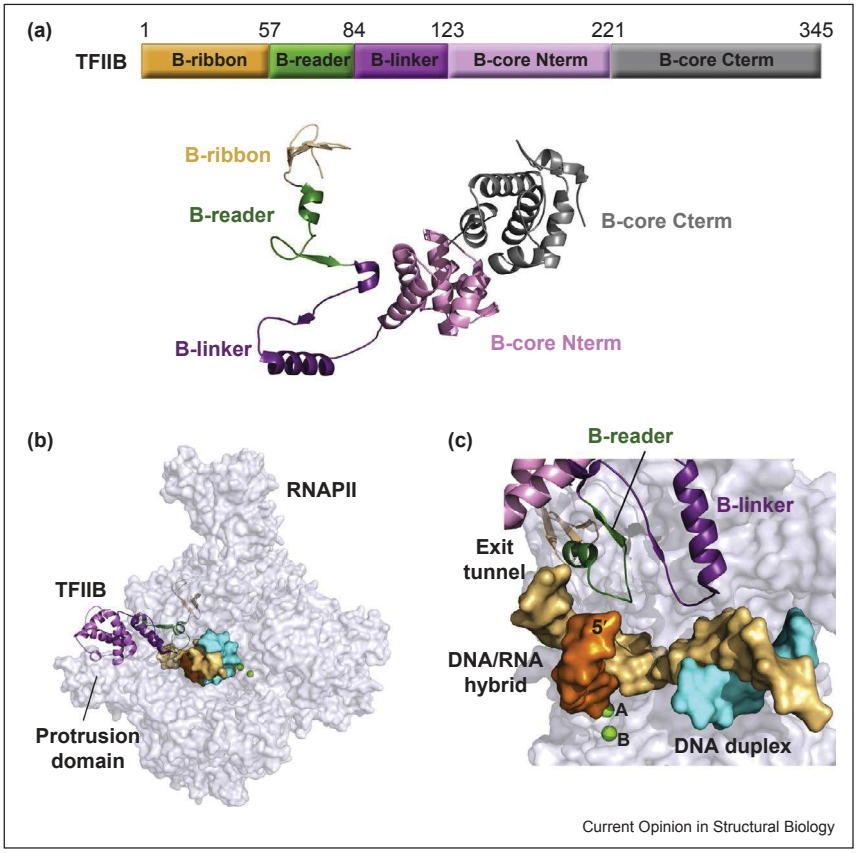
The recent 1.97 Å crystal structure of a yeast TAF1 TAND–TBP complex, together with mutational data, provides a complete atomic description of TAF1–TBP interactions [11•]. A single-chain TAF1 TAND–TBP construct was utilized to stabilize complex formation. Yeast TAF1 contains two TAND domains, TAND1 and TAND2, which bind to the concave and convex surfaces of TBP, respectively (Figure 1c). In transcription, TAND1 itself was shown to function as an activation domain, while TAND2 independently exerts an inhibitory effect [22]. The interactions observed at the TAND1–TBP binding interface are mainly hydrophobic and structurally mimic those observed in the complex formed by TBP and TATA DNA [23] (Figure 1c). On the other hand, negatively charged residues of TAND2 interact with positively charged residues on the convex surface of TBP, and a phenylalanine residue in TAND2 (F57) serves as a hydrophobic anchor of the TAND2–TBP interaction (Figure 1c). This TAND2 aromatic and acidic peptide region represents a conserved regulatory TBP binding motif, which the authors could identify also in other, structurally distinct transcription factors including TFIIA, Brf1 and Mot1, conveying competition between such binding domains as a means of regulation [11•]. The present TAF1–TBP crystal structure shows the simultaneous integration of activating and repressing transcriptional modalities in the complex, providing an atomic framework for further analysis.

RNAPII–TFIIB initially transcribing complex (ITC)

TFIIB contributes to core promoter binding and melting, transcription start site (TSS) selection and transcript separation by its N-terminal region interacting with DNA and RNAPII [24,25]. The structural basis of these activities remained elusive. Crystal structures of an RNAPII–TFIIB complex, and RNAPII–TFIIB with DNA template and RNA, reveal previously unobserved interactions of TFIIB with RNAPII and nucleic acids, shedding new light on important roles of TFIIB in transcription initiation [12••].

TFIIB consists of two globular cyclin folds and an extended N-terminal domain comprising the so-called B-ribbon, B-reader and B-linker elements (Figure 2a). During transcription initiation, TFIIB extends with its B-reader element into the RNAPII active center cleft. In earlier studies of RNAPII–TFIIB complexes, the B-reader could not be localized completely [7,26]. In the recent crystal structure of a RNAPII–TFIIB complex refined at 3.4 Å resolution, the complete B-reader could now be traced unambiguously (Figure 2a) [12••]. A truncated version of TFIIB lacking the C-terminal cyclin fold and a mobile N-teminal tail was used. TFIIB interacts intricately with RNAPII: with the dock (B-ribbon), the wall (B-core N-terminal cyclin fold), the RNA exit tunnel (B-reader helix) and the clamp (B-linker helix). Upon TFIIB binding, domains of RNAPII are rearranged, leading to a partial closure of the cleft. A mobile region at the tip of the Rpb2 protrusion domain in RNAPII adopts a defined structure, which is required for the stability of the initiation complex. The interactions between TFIIB and RNAPII explain functional studies,

Figure 2



The RNAPII–TFIIB initially transcribing complex (ITC). **(a)** The domain organization of TFIIB is drawn in a schematic fashion (top) with the structure of TFIIB derived from the RNAPII–TFIIB crystal structure [12••]. The C-terminal cyclin fold (B-core Cterm) is not present in the crystal and was modeled based on the TFIIB–TBP–TATA DNA complex structure (PDB code 1C9B). **(b)** The crystal structure of the ITC is shown, comprising RNAPII (colored in gray), TFIIB (domains color-coded as above), DNA template (strands colored in cyan and yellow), and a 6-nucleotide RNA strand (colored in orange). Binding of TFIIB to RNAPII stabilizes the protrusion domain (marked). **(c)** Zoom-in on the ITC active site. The B-reader (colored in green) contacts the DNA–RNA hybrid. Electrostatic repulsion by an acidic patch of the B-reader loop may assist in guiding the 5′ end of an emerging RNA transcript toward the exit tunnel of RNAPII. Two active site metal ions, denoted A and B, are shown as green spheres [12••].

which showed that a mutation in the B-reader domain of TFIIB that causes a shift in TSS, could be compensated for by mutations on RNAPII in the Rbp2 subunit [27].

In the absence of nucleotides, only one catalytic metal ion is present in the active site of RNAPII [28]. The RNAPII–TFIIB crystal structure reveals the presence of two metal ions (A and B), which are both required for the catalytic activity of RNAPII. Without directly contacting the active site, TFIIB induces rearrangements in RNAPII that increase the affinity for the second metal ion. Consistent with this, the authors show biochemically that addition of TFIIB stimulates the RNA chain initiation activity of RNAPII [12••].

A DNA template comprising a downstream DNA duplex, and a 6 nucleotide RNA strand were incorporated into the RNAPII–TFIIB complex, giving rise to the initially transcribing complex (ITC) (Figure 2b). The ITC structure suggests that, in addition to allosteric active-site rearrangements, TFIIB further stimulates RNA synthesis by its B-reader interacting directly with DNA (Figure 2b,c). The B-reader prevents the tilting of short DNA–RNA hybrids which would otherwise lead to premature abortion, assists in defining the TSS and may stabilize the melted DNA conformation in the open promoter complex [12••].

In addition, the structure of the ITC suggests a mechanism for RNA separation from DNA by the B-reader element. Acidic residues in the B-reader loop may repel the 5′-triphosphate of the emerging RNA strand, which, in conjunction with attraction of the 5′-triphosphate to a cluster of positively charged residues in the exit tunnel, would lead to separation of the RNA strand from the DNA in the RNAPII active site cleft. When the growing

RNA reaches a length of 12–13 nucleotides, steric clashes with B-reader and B-ribbon cause displacement of TFIIB from RNAPII, and the elongation complex is formed.

An educative movie compiles these and other key aspects of initiation, transition from initiation to elongation and elongation by RNAPII, providing a comprehensive view of our current understanding of RNAPII transcription [29••].

Architecture of human and yeast PICs revealed

Structural information on PIC architecture, containing RNAPII, GTFs and promoter DNA, has been elusive. Two recent landmark studies now provide first structural snapshots of the human and yeast PIC [13••,14••].

Human PIC assembly

In order to assemble human PIC *in vitro*, reconstitution experiments were performed in which PIC components were added sequentially to a core promoter DNA template derived from the SCP [15]. EM structures of these complexes were determined, revealing the location of subsequently added PIC components [13••]. A TBP–TFIIA–TFIIB–promoter DNA complex bound to RNAPII was first reconstituted and analyzed. Addition of TFIIF positions the downstream DNA along the RNAPII cleft, involving direct interactions of TFIIF with core promoter DNA. The RNAPII clamp domain opens as it accommodates the downstream DNA, and a rotation of the TBP–TFIIA–TFIIB subcomplex was observed. This rotation repositions TFIIB in close proximity to RNAPII. TFIIE addition stabilizes a closed promoter complex, in which TFIIE, TFIIB and TFIIF are localized in the vicinity of promoter DNA, suggesting structural cross-talk between these three GTFs for promoter opening [13••].

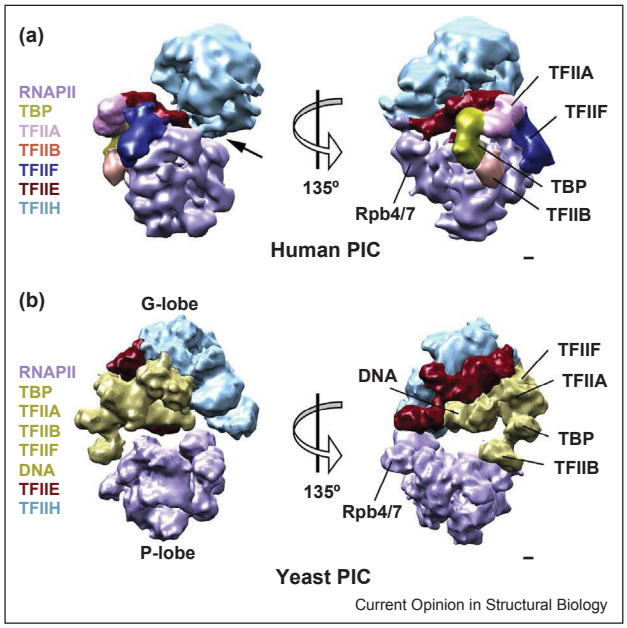
An open promoter complex (OC) was created by adding TFIIH to the closed promoter complex. Transition to the OC is catalyzed by a helicase subunit of TFIIH, XBP, and is assisted by TFIIE [30]. In the corresponding negative-stain EM reconstruction, TFIIH forms only few contacts with other components of the PIC. Particularly, it interacts with the stalk domain of RNAPII at the TFIIE interaction site (Figure 3a). TFIIH is thought to unwind the template DNA by an unconventional 'screw-in' mechanism, with a torque being generated by rotation of the downstream DNA [31]. The position of XPB at +10 to +20 base pairs from the TSS, deduced from the EM structure, supports this 'screw-in' DNA translocase activity of TFIIH [13••]. The human PIC structures and related biochemical studies recently have been reviewed in more detail elsewhere [32].

Architecture of a yeast PIC

In a parallel study, a yeast PIC was characterized by using cryo-EM (Figure 3b) [14••]. As in the human PIC analysis, TBP was used instead of holo-TFIID. PICs were assembled by incubating RNAPII, TBP and the GTFs TFIIA, TFIIB, TFIIE, TFIIF, TFIIH and TFIIS from *Saccharomyces cerevisiae* with a fragment of *HIS4* promoter DNA (−81/+1), followed by glycerol gradient centrifugation. The locations of all subunits were assigned in EM reconstructions with the help of protein–protein interaction data derived from cross-linking and mass spectrometry experiments.

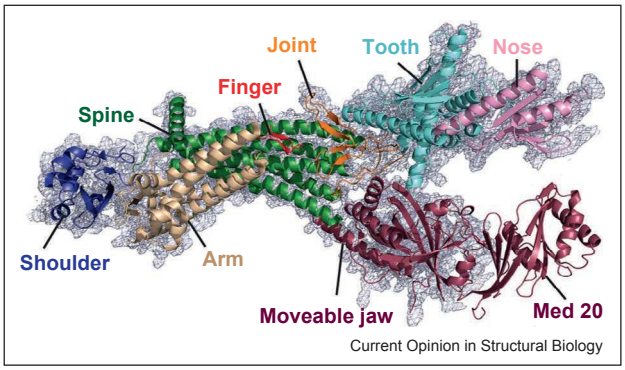
The yeast PIC adopts a compact shape containing two distinct lobes, a P-lobe comprising RNAPII and a G-lobe harboring the GTFs (Figure 3b). The promoter DNA is associated only with TFIIB, TBP and TFIIE in the G-lobe at its upstream end, and with TFIIH at its downstream end. The DNA is not in contact with RNAPII, but rather hovering above the active site cleft (Figure 3b). In

Figure 3



EM structures of human and yeast PIC assemblies containing RNAPII and GTFs. **(a)** Two views of a human PIC structure are shown, related by a 135° rotation. RNAPII (colored in violet) directly interacts with the GTFs. Locations of GTFs are marked. TFIIH (colored in cyan) only associates loosely with the PIC, primarily involving interactions with the RNAPII stalk domain (marked by an arrow). The GTFs are oriented toward RNAPII in a way that the promoter DNA (not localized in this structure) is inserted into RNAPII, descending into the cleft by means of a 'screw-in' mechanism catalyzed by the helicase activity of TFIIH [13••]. **(b)** The structure of a yeast PIC assembly is depicted in the same orientations as human PIC shown in (a). The RNAPII subunits Rpb4/7 (marked) were used to orient the structures. Yeast PIC exhibits two distinct lobes. The P-lobe contains RNAPII (colored in violet). The G-lobe contains GTFs and DNA. The entire density corresponding to GTFs, TBP, TFIIA and TFIIB and DNA are colored in yellow. The promoter DNA is associated with the GTFs in the G-lobe and positioned above the RNAPII cleft [14••]. Scale bars, 1 nm.

Figure 4



Crystal structure of *S. pombe* Mediator head. The structure resembles the head and limb of a crocodile, comprising 8 distinct structural elements (shown as ribbons), superimposed on the experimental electron density map at 3.4 Å resolution (blue mesh, contoured at 1σ). The Med20 subunit (labeled), which was not present in the crystal structure, was placed by modeling [36•].

this orientation, the promoter DNA interacts with the Ssl2 subunit of TFIIH, which is the yeast homolog of XPB helicase [14••]. Yeast PIC appears to be captured in a state ready for promoter opening, and the descent of the DNA into the RNAPII cleft for transcription initiation. The yeast PIC EM structure, refined at 16 Å resolution accounts for most of the protein mass present in this PIC assembly, and provides a wealth of molecular level information on PIC architecture in unparalleled detail [14••].

Crystal structures of Mediator head module

Mediator is a megadalton-sized central regulator of eukaryotic transcription and plays a key role in connecting gene-specific regulatory factors with RNAPII [33]. Mediator is made up of ~25 subunits in *S. cerevisiae*, and comprises three distinct structural modules, called head, middle and tail. The head module directly contacts RNAPII. Three crystal structures of the Mediator head from baker's and fission yeast were elucidated [34,35,36•], showing an overall shape resembling a molecular wrench [34] and illustrating how the head module interacts with the C-terminal domain (CTD) of RNAPII [35]. Mediator head from *S. pombe* yielded particularly well-ordered crystals, diffracting X-rays to higher resolution (3.4 Å) as compared to the complexes from *S. cerevisiae*, providing the most complete Mediator head module structure to date and revealing its intricate architecture (Figure 4) [36•].

Conclusions

Here we have reviewed recent advances in our understanding of RNAPII transcription initiation, derived from recently determined crystal structures and EM analyses. Improvements in recombinant expression techniques, *in vitro* reconstitution approaches and powerful integrated structure determination methods have been instrumental to bring about these major steps forward. We now have first structural impressions of very large assemblies of the entire PIC, containing more than 30 proteins and promoter DNA. We anticipate that further studies of PICs containing complete TFIID and Mediator complexes, will be forthcoming, complemented by atomic studies revealing the inner workings of RNAPII and the GTFs, and their interaction with the DNA template in ever increasing detail, toward resolving the molecular puzzle of the eukaryotic transcription initiation process.

Acknowledgements

We thank all members of the Berger and Schaffitzel laboratories for helpful discussions. EK and ST acknowledge support from a Marie-Curie Intra-European Fellowship (IEF) by the European Commission (EC), Framework Programme (FP) 7. MH is recipient of a Kekulé predoctoral fellowship by the Fonds der Chemischen Industrie (FCI, Germany). IB is supported by an EC FP7 collaborative project (CP), ComplexINC (grant number 279039).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Morse RH: **Transcription factor access to promoter elements**. *J Cell Biochem* 2007, **102**:560-570.

2. Gill G: **Regulation of the initiation of eukaryotic transcription**. *Essays Biochem* 2001, **37**:33-43.

3. Thomas MC, Chiang CM: **The general transcription machinery and general cofactors**. *Crit Rev Biochem Mol Biol* 2006, **41**:105-178.

4. Hoiby T, Zhou H, Mitsiou DJ, Stunnenberg HG: **A facelift for the general transcription factor TFIIA**. *Biochim Biophys Acta* 2007, **1769**:429-436.

5. Papai G, Weil PA, Schultz P: **New insights into the function of transcription factor TFIID from recent structural studies**. *Curr Opin Genet Dev* 2011, **21**:219-224.

6. Vannini A, Cramer P: **Conservation between the RNA polymerase I, II, and III transcription initiation machineries**. *Mol Cell* 2012, **45**:439-446.

7. Kostrewa D, Zeller ME, Armache KJ, Seizl M, Leike K, Thomm M, Cramer P: **RNA polymerase II–TFIIB structure and mechanism of transcription initiation**. *Nature* 2009, **462**:323-330.

8. Luse DS: **Rethinking the role of TFIIF in transcript initiation by RNA polymerase II**. *Transcription* 2012, **3**:156-159.

9. • Cianfrocco MA, Kassavetis GA, Grob P, Fang J, Juven-Gershon T, Kadonaga JT, Nogales E: **Human TFIID binds to core promoter DNA in a reorganized structural state**. *Cell* 2013, **152**:120-131.
Cryo-EM structures of human TFIID, purified from native source, are provided including a complex of TFIID with promoter DNA and TFIIA. TFIID is shown to exist in two distinct forms, a canonical and a reorganized form. The reorganized form is stabilized by DNA and TFIIA binding.

10. • Bieniossek C, Papai G, Schaffitzel C, Garzoni F, Chaillet M, Scheer E, Papadopoulos P, Tora L, Schultz P, Berger I: **The architecture of human general transcription factor TFIID core complex**. *Nature* 2013, **493**:699-702.
Cryo-EM structure of subassemblies of recombinant human TFIID, including a presumed physiological core–TFIID complex containing two copies each of TAF4, 5, 6, 9 and 12. This core–TFIID has twofold symmetry. Accretion of one copy each of two further TAFs, TAF8 and TAF10, to core–TFIID results in large structural rearrangements and an asymmetric complex. A pathway for the assembly of holo-TFIID is suggested.

11. • Anandapadamanaban M, Andresen C, Helander S, Ohyama Y, Siponen MI, Lundstrom P, Kokubo T, Ikura M, Moche M, Sunnerhagen M: **High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation**. *Nat Struct Mol Biol* 2013, **20**:1008-1014.
Crystal structure of yeast TBP bound to the yeast TAF1 N-terminal domains, TAND1 and TAND2. The structural basis for the activating and repressing functions of TAF1 on transcription initiation, mediated by the TAND1 and TAND2 domains binding to the concave and convex surfaces of TBP, respectively, is revealed at 1.97 Å resolution.

12. •• **Sainsbury S, Niesser J, Cramer P: Structure and function of the initially transcribing RNA polymerase II–TFIIB complex**. *Nature* 2013, **493**:437-440.
Crystal structures of RNAPII–TFIIB and an initially transcribing complex (ITC) containing a DNA–RNA hybrid. The structure of the N-terminal region of TFIIB including the entire B-reader, is revealed. A structural basis for TFIIB stabilizing the melted DNA template, and directing the emerging RNA transcript to the exit tunnel is provided.

13. •• **He Y, Fang J, Taatjes DJ, Nogales E: Structural visualization of key steps in human transcription initiation**. *Nature* 2013, **495**:481-486.
EM reconstructions of step-wise assembled intermediates of human PIC, are described, providing important clues of PIC assembly and promoter opening. TBP, promoter DNA, TFIIA, TFIIB, TFIIF and RNAPII were reconstituted. Subsequent addition of TFIIE stabilizes this PIC core assembly by interaction with RNAPII and TFIIF, which serves as a scaffold for TFIIH binding. The XPB helicase subunit of TFIIH is positioned to induce promoter opening by a 'screw-in' mechanism, pushing the DNA into the RNAPII active site for RNA synthesis.

14. •• Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalisman Y, Liu X, Levitt M *et al.*: **Architecture of an RNA polymerase II transcription pre-initiation complex**. *Science* 2013, **342** http://dx.doi.org/10.1126/science.1238724.
This structural study of a large PIC assembly from yeast, comprising 32 polypeptides, combines cryo-EM with chemical cross-linking and mass spectrometry (xl-MS). Two distinct lobes are described, a P-lobe containing RNAPII and a G-lobe containing TBP, promoter DNA and GTFs TFIIA, TFIIB, TFIIE, TFIIF and TFIIH. In this structure, the promoter DNA is only in contact with the G-lobe and not with RNAPII.

15. Juven-Gershon T, Cheng S, Kadonaga JT: **Rational design of a super core promoter that enhances gene expression**. *Nat Methods* 2006, **3**:917-922.

16. Cianfrocco MA, Nogales E: **Regulatory interplay between TFIID's conformational transitions and its modular interaction with core promoter DNA**. *Transcription* 2013, **4**:1-7.

17. Sanders SL, Garbett KA, Weil PA: **Molecular characterization of *Saccharomyces cerevisiae* TFIID**. *Mol Cell Biol* 2002, **22**:6000-6013.

18. Wright KJ, Marr MT 2nd, Tjian R: **TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter**. *Proc Natl Acad Sci U S A* 2006, **103**:12347-12352.

19. Gangloff YG, Romier C, Thuault S, Werten S, Davidson I: **The histone fold is a key structural motif of transcription factor TFIID**. *Trends Biochem Sci* 2001, **26**:250-257.

20. Kotani T, Miyake T, Tsukihashi Y, Hinnebusch AG, Nakatani Y, Kawaichi M, Kokubo T: **Identification of highly conserved amino-terminal segments of dTAFII230 and yTAFII145 that are functionally interchangeable for inhibiting TBP–DNA interactions in vitro and in promoting yeast cell growth in vivo**. *J Biol Chem* 1998, **273**:32254-32264.

21. Liu D, Ishima R, Tong KI, Bagby S, Kokubo T, Muhandiram DR, Kay LE, Nakatani Y, Ikura M: **Solution structure of a TBP–TAF(II)230 complex: protein mimicry of the minor groove surface of the TATA box unwound by TBP**. *Cell* 1998, **94**:573-583.

22. Kotani T, Banno K, Ikura M, Hinnebusch AG, Nakatani Y, Kawaichi M, Kokubo T: **A role of transcriptional activators as antirepressors for the autoinhibitory activity of TATA box binding of transcription factor IID**. *Proc Natl Acad Sci U S A* 2000, **97**:7178-7183.

23. Kim Y, Geiger JH, Hahn S, Sigler PB: **Crystal structure of a yeast TBP/TATA-box complex**. *Nature* 1993, **365**:512-520.

24. Knutson BA, Hahn S: **TFIIB-related factors in RNA polymerase I transcription**. *Biochim Biophys Acta* 2013, **1829**:265-273.

25. Liu X, Bushnell DA, Kornberg RD: **RNA polymerase II transcription: structure and mechanism**. *Biochim Biophys Acta* 2013, **1829**:2-8.

26. Liu X, Bushnell DA, Wang D, Calero G, Kornberg RD: **Structure of an RNA polymerase II–TFIIB complex and the transcription initiation mechanism**. *Science* 2010, **327**:206-209.

27. Chen BS, Hampsey M: **Functional interaction between TFIIB and the Rpb2 subunit of RNA polymerase II: implications for the mechanism of transcription initiation**. *Mol Cell Biol* 2004, **24**:3983-3991.

28. Cramer P, Bushnell DA, Kornberg RD: **Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution**. *Science* 2001, **292**:1863-1876.

29. •• Cheung AC, Cramer P: **A movie of RNA polymerase II transcription**. *Cell* 2012, **149**:1431-1437.
A movie capturing central structural aspects of RNAPII transcription initiation, initiation-elongation transition and elongation was created by combining available structures determined by X-ray crystallography and electron microscopy, models derived from fluorescence resonance energy transfer (FRET) and protein crosslinking studies. This illustrative teaching tool can be downloaded from http://www.sciencedirect.com/science/article/pii/S0092867412006976.

30. Grunberg S, Warfield L, Hahn S: **Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening**. *Nat Struct Mol Biol* 2012, **19**:788-796.

31. Egly JM, Coin F: **A history of TFIIH: two decades of molecular biology on a pivotal transcription/repair factor**. *DNA Repair (Amst)* 2011, **10**:714-721.

32. Grunberg S, Hahn S: **Structural insights into transcription initiation by RNA polymerase II**. *Trends Biochem Sci* 2013, **38(12)**:603-611.

33. Poss ZC, Ebmeier CC, Taatjes DJ: **The Mediator complex and transcription regulation**. *Crit Rev Biochem Mol Biol* 2013, **48(6)**:575-608.

34. Imasaki T, Calero G, Cai G, Tsai KL, Yamada K, Cardelli F, Erdjument-Bromage H, Tempst P, Berger I, Kornberg GL *et al.*: **Architecture of the Mediator head module**. *Nature* 2011, **475**:240-243.

35. Robinson PJ, Bushnell DA, Trnka MJ, Burlingame AL, Kornberg RD: **Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II**. *Proc Natl Acad Sci U S A* 2012, **109**:17931-17935.

36. • Lariviere L, Plaschka C, Seizl M, Wenzeck L, Kurth F, Cramer P: **Structure of the Mediator head module**. *Nature* 2012, **492**:448-451.
Crystal structure of Mediator head module from *S. pombe*, determined at 3.4 Å resolution, provides a close to complete view of Mediator head architecture resembling a 'crocodile head'. The extended architecture of the head module suggests numerous interactions not only with RNAPII but also with other components of the transcription initiation machinery.

2. MATERIALS AND METHODS

SUMMARY IN ENGLISH

In this chapter, the materials and methods are described which were used in the work presented here. Detailed description of all methods used for cloning, protein production (in insect cells and bacterial cells) and the purification of proteins/protein complexes is provided. Other biophysical and biochemical methods are also described, including native MS, EMSA, analytical SEC and pulldowns, AUC, CLMS, peptide array, mass spectrometry, SEC-MALLS etc. All the methods used for structural characterization of different TFIID subassemblies are explained in much more details which include protein crystallization and structure determination by X-ray crystallography, NMR analysis and structure determination, SAXS analysis and modeling, negative stain EM analysis and RCT model calculations.

RÉSUMÉ EN FRANÇAISE

Dans ce chapitre, les matériel et méthodes qui ont été utilisés au cours de ce travail de thèse sont décrits. Commençant par les détaille sur les méthodes utilisées pour le clonage, l'expression (en cellules d'insectes et en cellules bactériennes) et la purification de protéines et de complexes protéiques. Des méthodes biophysiques et biochimiques, dont la MS natif, EMSA, SEC analytique et les pull-downs, AUC, CLMS, réseau de peptides, la spectrométrie de masse, la SEC-MALLS sont aussi expliqué. Toutes les méthodes utilisées pour la caractérisation structurale des différents sous-ensembles de TFIID sont expliqués de façon plus détaillés. Ces dernières sont les stratégies de cristallisation de protéines et de détermination de la structure par cristallographie aux rayons X, l'analyse par RMN et la détermination de la structure, l'analyse et la modélisation SAXS, analyse d'EM et de calculs de modèles RCT.

2.1. Materials

2.1.1. Chemicals, Enzymes, Consumables and Equipments

All chemicals, enzymes, consumables and equipments used in this work are listed in Table 4, Table 5, Table 6 Table 7. All chemicals were of analytical grade, if not otherwise stated. Other than the equipments listed in Table 6, instrumentation used at different data collections facilities for X-ray diffraction, SAXS, NMR and EM data collection are mentioned in methods or results.

Table 4: Chemicals.

Chemical	Source
2-Mercaptoethanol (β-ME)	Euromedex
2-Propanol	Euromedex
4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES)	Euromedex
2-(N-morpholino)ethanesulfonic acid (MES)	Sigma-Aldrich
Acetate, sodium salt (Na-Acetate)	Sigma-Aldrich
Acetic acid	Euromedex
Acrylamide/bis-acrylamide, 30% (37.5:1)	Euromedex
Acrylamide/bis-acrylamide, 40% (29:1)	Roth
Agar-Agar	Euromedex
Agarose, electrophoresis grade	Euromedex
Ammonium Acetate (NH_4-Acetate)	Merck
Ammonium bicarbonate	Fluka
Ammonium Chloride	ICN biomedicals
Ammonium Chloride (^{15}Nlabeled)	Eurisotop
Ammonium peroxodisulfate (APS)	Sigma-Aldrich
Ampicillin (Amp)	Sigma-Aldrich
BCIP/NBT stock solution	Roche
Betaine solution 5M	Sigma-Aldrich
Bluo-Gal	Invitrogen
Boric acid	Euromedex
Bromophenol blue	Euromedex
BS3(Sulfo-DSS)	Thermo
BSA, 100x (10 mg/mL)	New England Biolabs
Calcium Chloride	Sigma-Aldrich
Chloramphenicol (Camp)	Sigma-Aldrich
Citrate, tri-sodium salt (Na-Citrate)	Sigma-Aldrich
cOmplete EDTA-free Protease inhibitor tablets	Roche

Coomassie Brilliant Blue G-250	Euromedex
Deoxynucleotides (dNTPs)	New England Biolabs
Dithiothreitol (DTT)	Euromedex
Dimethylformamide (DMF)	Sigma-Aldrich
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich
DNA molecular weight marker (1KB, 100bp)	New England Biolabs
DNA molecular weight marker, EZ load (20bp)	Bio-rad
Ethanol	Euromedex
Ethidium bromide solution (10 mg/mL)	Euromedex
Ethylene Glycol	Sigma-Aldrich
Ethylenediaminetertaacetate, disodium salt (EDTA)	Sigma-Aldrich
Formate, sodium salt (Na-Formate)	Sigma-Aldrich
Gentamycin (Gent)	Sigma-Aldrich
Glucose	Euromedex
Glucose (^{13}C labeled)	Eurisotop
Glutaraldehyde 25% solution	Electron Microscopy Sciences
Glycerol	Euromedex
Glycine	Euromedex
Hydrochloric acid (HCl)	Euromedex
Imidazole	Sigma-Aldrich
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Euromedex
Kanamycin (Kan)	Sigma-Aldrich
LB medium powder	Euromedex
Leupeptin hemisulphate	Sigma-Aldrich
L-Lysine	Sigma-Aldrich
Lysozyme	Euromedex
Magnesium Chloride ($MgCl_2$)	Sigma-Aldrich
Magnesium Sulphate heptahydrate	Sigma-Aldrich
Malonate, sodium salt (Na-Malonate)	Sigma-Aldrich
Maltose monohydrate D (+)	Sigma-Aldrich
MEM vitamins 100X	Gibco (Life technologies)
Methanol	VWR
N,N,N',N'-Tetramethylethylenediamine (TEMED)	Euromedex
Nu-PAGE 4-12% bis tris gel	Novex
Pepstatin A	Sigma-Aldrich
Poly-ethylene glycol (PEG) 200-20000	Hampton Research Inc.
Potassium chloride	Euromedex
Potassium phosphate monobasic anhydrous	Euromedex
Prestained protein molecular weight marker (10-250 kDa)	Bio-rad
Sf-900™ III SFM Insect cell medium	Gibco (Life technologies)
Skimmed milk powder	Roth
Sodium Acetate (Na-Acetate)	Sigma-Aldrich
Sodium chloride (NaCl)	Euromedex
Sodium dodecyl sulfate (SDS)	Euromedex
Sodium hydroxide (NaOH)	Euromedex
Sodium Fluoride	Sigma-Aldrich

Sodium phosphate dibasic	Sigma-Aldrich
Sodium/potassium tartrate, (NH_4/K- tartrate)	Sigma-Aldrich
Sucrose	Euromedex
Tetracycline (Tet)	Sigma-Aldrich
Tricine	Sigma-Aldrich
Tris(hydroxymethyl)aminomethane (Tris)	Euromedex
Tween-20	Sigma-Aldrich
Uranyl Acetate	Electron Microscopy Sciences
X-tremeGENE DNA transfection reagent	Roche
Xylene cyanol FF	Sigma-Aldrich

Table 5: Enzymes and Antibodies.

Enzyme	Source
Benzonase	Invitrogen
Cre-Recombinase	EMBL
Phusion® High-Fidelity DNA Polymerase	New England Biolabs
Restriction enzymes	New England Biolabs
T4 DNA Polymerase	New England Biolabs
TEV Protease	EMBL
T4 DNA Ligase	New England Biolabs
Trypsin	Sigma-Aldrich
Anti-mouse HRP conjugated secondary antibody	Sigma-Aldrich
Mouse monoclonal anti-Histidine AP conjugated antibody	Sigma-Aldrich
Mouse monoclonal anti-Histidine primary antibody	Sigma-Aldrich

Table 6: Equipment and Consumables.

Material	Source
ÄKTA FPLC systems (Prime, Basic, Purifier, Micro)	GE Healthcare
Amicon Ultra spin concentrators (3kDa-50kDa, 0.5 mL-15 mL)	Millipore
Avanti J-26XP	Beckman Coulter
Centrifuge 5424	Eppendorf
Centrifuge 5804	Eppendorf
Centrifuge 5427	Eppendorf
Chemidoc MP	Bio-rad
Cover slides; round, 22 mm, siliconized	Hampton Research
Crystallization plates; 24-well; hanging drop and sitting drop	Hampton Research
Dialysis membrane (3000 to 8000 MW cutoff)	Sigma Aldrich
Dialysis cassettes	Sigma Aldrich
Eppendorf tubes; 1.5 mL / 2.0 mL	Eppendorf
Falcon tubes; 15 mL/50 mL	BD Biosciences

Gravity flow columns	Bio-rad/Thermo Scientific
Incubator Multitron	Infors HT
PVDF membrane	Millipore
pH Meter	InoLab
Pipettes (2-1000 µL)	Gilson/Eppendorf
Rotors (JA25.50, JA8.100, JA20.0, Ti70, SW32.0)	Beckman Coulter
Spectrophotometer Nanodrop 1000	Thermo Scientific
Thermocycler T3000	Biometra
Western blot semidry apparatus	Biorad
Whatman filter paper	Bia-rad
XK16 Column	GE healthcare

Table 7: Commercial kits.

Name	Purpose	Source
QIAprep Spin Miniprep Kit	Plasmid extraction and purification	Qiagen
QIAquick PCR Purification Kit	PCR product purification	Qiagen
QIAquick Gel Extraction Kit	DNA extraction from agarose gels	Qiagen
ECl western blot kit	Detection of HRP in western blot	Pierce (Thermo Scientific)

2.1.2. Primers

All oligonucleotides used in this study were synthesized by Life Technologies and supplied in desalted quality at a 25 nmol synthesis scale. Oligonucleotides were dissolved in ultrapure water and stored at -20°C until use. Table 8 lists all oligonucleotides used in this work with their corresponding sequence.

Table 8: Primers.

FWD: forward, REV: reverse, SLIC: Sequence and Ligation Independent Cloning. The cloning method or restriction enzyme used is indicated.

Name	Description	Sequence 5'→ 3'
KG2	pFB_TAF11^{201}, REV, EcoRI	GCAATCGAATTCCTAAGGGATCTGTCCTTTTGACTTTAACC
KG3	pFB_TAF11^{211},	GCAATCGAATTCCGGATCCCTAGAAGAAG

	REV, EcoRI	
KG8	pFB_TAF13[1], FWD, XmaI	GATGTTCCCGGGATGGCAGATGAGGAAGAAGACC
KG9	pU_TAF1[26], FWD, SLIC	CTGTATTTTCAGGGCAGCGACGAAGATTCCGCTGGAGGCGGCCC
KG10	pU_TAF1[87], REV, SLIC	GAAAGCGGCCGCTAGCTCACAATTCTTCATTTGCCGTGAGTTCAGT GATC
KG11	pU_backbone TAF1[26-87], FWD, SLIC	GCAAATGAAGAATTGTGAGCTAGCGGCCGCTTTCGAATCTAGAGC CTGC
KG12	pU_backbone TAF1[26-87/168], REV, SLIC	GGAATCTTCGTCGCTGCCCTGAAAATACAGGTTTTCAGTCTGCGC
KG13	pU_TAF1[168], REV, SLIC	GAAAGCGGCCGCTAGCTCACTTCTTCATTGGTCCCGGGGGTGGAG GTGG
KG14	pU_backbone TAF1[26-168], FWD, SLIC	GGACCAATGAAGAAGTGAGCTAGCGGCCGCTTTCGAATCTAGAGC CTGC
KG35	pFB_TAF11[87], FWD, NcoI	GAGTGCCCATGGGAGATACCAAAGAAAAGAAAGAGAAAAAGC
KG48	pFB_TAF11[21], FWD, NcoI	GATGTTCCATGGGAACGGCCGCTGTGCCCGGGGAC
KG49	pFB_TAF11[41], FWD, NcoI	GATGTTCCATGGGAGACGGAGACGCAGATGTGGACTTG
KG50	pFB_TAF11[61], FWD, NcoI	GATGTTCCATGGGACAGGATGTCTCAGATTTAACAACAG
KG51	pFB_TAF11[51], FWD, NcoI	GATGTTCCATGGGAGCTGCAGCGGAGGAAGGCGAG
KG52	pFB_TAF13[97], REV, NsiI	GCAATCATGCATCTAAAACTTCCTTGGGTCCTTTCG
KG53	pFB_TAF13[28], FWD, XmaI	GATGTTCCCGGGATGAGAAAGAGACTTTTTTCTAAAGAATTGCG
KG54	pFB_TAF13[124], REV, NsiI	GCAATCATGCATTCAAGATCCATAATTTGCTTCATC
KG64	pET28(a)_TBP[FL], FWD, NdeI	GATGTTCATATGGGTGAAAACCTGTATTTTCAGGGC
KG65	pET28(a)_TBP[FL], REV, KpnI	GCAATCGGTACCTTACGTCGTCTTCCTGAATCCCTTTAG
KG70	pMAL_GB1, FWD, SLIC	CAAGGACCATAGATTATGAAACAGTACAAGCTGATTCTGAACGG
KG71	pMAL_GB1, REV, SLIC	GTGGCAGAAGAGTTCATATGACCCTGGAAGTACAGATTTTCGC
KG72	pMAL_TAF1[1157-1207]_backbone	CTTCCAGGGTCATATGAACTCTTCTGCCACTGGACGCTGTCTCAAG

	(GB1), FWD, SLIC	
KG73	pMAL_TAF1$^{1157-1207}$_backbone (GB1), REV, SLIC	CTTGTACTGTTTCATAATCTATGGTCCTTGTTGGTGAAGTGCTC
KG80	pMAL_Trx, FWD, SLIC	CAAGGACCATAGATTATGAGCGATAAAATTATTCACCTGACTGAC
KG81	pMAL_Trx, REV, SLIC	AGAAGAGTTGCTAGCGGCCAGGTTAGCGTCGAGGAACTCTTTCAAC
KG82	pMAL_TAF1$^{1157-1207}$_backbone (Trx), FWD, SLIC	GACGCTAACCTGGCCGCTAGCAACTCTTCTGCCACTGGACGCTGTC
KG83	pMAL_TAF1$^{1157-1207}$_backbone (Trx), REV, SLIC	AATTTTATCGCTCATAATCTATGGTCCTTGTTGGTGAAGTGCTCG
KG131	pET28(a)_HisMBP, FWD, NdeI	GACTAACATATGAAAACTGAAGAAGGTAAACTGG
KG132	pET28(a)_HisMBP, REV, BamHI	GATTTAGGATCCTTATCACGAGCTCGAATTAGTCTG
KG143	pIDC_TAF1^{578},FWD, BamHI	GATGTTGGATCCATGCAGGGTCTTCGAGGCACCTTTG
KG144	pIDC_TAF1^{1210}, REV, RsrII	GCAATCCGGTCCGTCATTAGGCAAATTTTCGAATGAATTCCTC
KG147	pDC_TAF1^{578}, FWD, NdeI	GATGTTCATATGCAGGGTCTTCGAGGCACCTTTG
KG148	pDC_TAF1^{1210}, REV, BamHI	GCAATCGGATCCTCATTAGGCAAATTTTCGAATGAATTCCTC
KG151	pIDC_TAF1^{1106}, REV, RsrII	GCAATCCGGTCCGTCATTAGTTCTCAATGTTCTTTCCCATTTC
KG153	pDC_TAF1^{1106}, REV, BamHI	GCAATCGGATCCTCATTAGTTCTCAATGTTCTTTCCCATTTC
KG159	pACE_TAF7^{1}, FWD, NdeI	GATGTTCATATGAGTAAAAGCAAAGATGATGCTCC
KG160	pACE_TAF7^{193},REV, BamHI	GCAATCGGATCCTCATTATTCTGCCTCCTTTGTTTCATCTTC
KG161	pFL_TAF7^{1}, FWD, SalI	GATGTTGTCGACATGAGTAAAAGCAAAGATGATGCTCC
KG162	pFL_TAF7^{193}, REV, NdeI	GCAATCCATATGTCATTATTCTGCCTCCTTTGTTTCATCTTC

2.1.3. Plasmids

All plasmids used in this study are listed in Table 9. The DNA sequence of all plasmids was verified by sequencing. All genes of interests were cloned as described in Chapter 2.2.2.

Table 9: Plasmids.

Name	Resistance	Source
pIDC_TAF1_578-1210	Cm	This Study
pIDC_TAF1_578-1106	Cm	This Study
pDC_TAF1_578-1210	Cm	This Study
pDC_TAF1_578-1106	Cm	This Study
pFL_10xHis_TEV_TAF7_1-193	Amp, Gent	This Study
pACE_10xHis_TEV_TAF7_1-193	Amp	This Study
pET28(a)_10xHis_TEV_TAF7_126-202	Kan	This Study, Berger group
pET28(a)_10xHis_TEV_TAF7_153-190	Kan	This Study
pET28(a)_10xHis_TEV_TAF7_156-190	Kan	This Study
pMAL_MBP_TEV_TAF1_1157-1207_6xHis	Amp	This Study, Berger group
pMAL_GB1_TEV_TAF1_1157-1207_6xHis	Amp	This Study
pMAL_Trx_TAF1_1157-1207_6xHis	Amp	This Study
pET28(a)_6xHis_MBP	Kan	This Study
pET28(a)_6xHis_TBPcore	Kan	Berger group
pET28(a)_6xHis_TEV_TBP-FL	Kan	This Study
pET28(a)_6xHis_scTFIIA	Kan	Berger group
pFB_6xHis_TEV_TAF11_TAF13	Amp	Berger group
pFB_6xHis_TEV_TAF11_87-211_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_21-211_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_41-211_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_61-211_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_51-211_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_41-201_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_51-201_TAF13	Amp	This Study
pFB_6xHis_TEV_TAF11_51-201_TAF13_1-97	Amp	This Study
pFB_6xHis_TEV_TAF11_51-201_TAF13_28-97	Amp	This Study
pFB_6xHis_TEV_TAF11_51-201_TAF13_28-124	Amp	This Study
pU_CBP_Strep_MBP_TAF1_26-87_6xHis	Cm	This Study
pU_CBP_Strep_MBP_TAF1_26-168_6xHis	Cm	This Study
pK_Dummy	Kan, Gent	Berger group

2.1.4. Bacterial strains

Different *E.coli* strains were used for cloning and expression. OmniMAX (Invitrogen) or Top10 (Invitrogen) cells were used for amplification of Acceptor plasmids or Acceptor-Donor fusions and also for amplification of pET28(a) or pMAL plasmids. Rosetta (DE3) (Novagen) or BL21 (DE3) star (Invitrogen) cells were used for bacterial expression. BW23473 cells were used for amplification of Donor plasmids. DH10-EMBacY cells were used for isolation of recombinant EMBacY bacmid.

2.1.5. Insect Cell lines

The *Spodoptera frugiperda* (*Sf*) insect cell line obtained from Invitrogen was used for expression using baculovirus based MultiBac system.

2.1.6. Chromatography Resins and Columns for protein purification

Talon resin (Clontech) or Ni^{2+}-NTA resin (Qiagen) were used for immobilized metal affinity chromatography (IMAC) purification of His-tagged proteins in gravity flow columns or XK16 columns using ÄKTA prime. Amylose resin (New England Biolabs) was used for pulldown assays. Chromatography columns used for protein purification in this study includes ion exchange columns (MonoQ 5/50GL, MonoS 5/50GL, HiTrapQ, HiTrap Heparin, HiTrap SPFF) and Size exclusion columns (SuperdexS75 10/300, SuperdexS75 10/300, SuperdexS75 16/60, SuperdexS200 16/60, SuperdexS200 PC3.2/30. All these columns were obtained from GE Healthcare and were used on ÄKTA Basic, Purifier or Micro system.

2.1.7. Crystallization screens, materials and reagents

The high-throughput crystallization (HTX) laboratory at EMBL Grenoble supplied all commercial crystallization screens. Other crystallization grade reagents and crystal handling tools were obtained from Hampton Research Inc. or Molecular Dimensions, if not otherwise stated.

2.1.8. Bioinformatics and computational tools and webservers

2.1.8.1. *In silico* DNA manipulation

APE: To visualize, translate and modify DNA sequences *in silico.* (http://biologylabs.utah.edu/jorgensen/wayned/ape/)

Cre-ACEMBLER: To generate Cre-LoxP fusion plasmids from individual Acceptor and Donor plasmids. (http://www.embl.fr/multibac/multiexpression_technologies/cre-acembler/)

2.1.8.2. Multiple sequence alignments

ClustalOmega (http://www.ebi.ac.uk/Tools/msa/clustalo/): To align protein sequences (McWilliam et al., 2013; Sievers and Higgins, 2014).

ESPRIPT (http://espript.ibcp.fr/ESPript/ESPript/): To visualize ClustalOmega results (Gouet et al., 1999).

2.1.8.3. Protein parameters, secondary structure and disorder prediction

ProtParam (http://web.expasy.org/protparam/): To calculate various physical and chemical parameters from protein sequence.

PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred): To predict secondary structure of proteins (Jones, 1999).

IUPRED (http://iupred.enzim.hu): For protein disorder prediction (Dosztanyi et al., 2005).

2.1.8.4. Crystal structure determination

XDS software package: To process crystallographic data (Kabsch, 2010).

CCP4 software package: For molecular replacement (Winn et al., 2011).

PHENIX software package: For structure refinement and modelling of molecular replacement template (Adams et al., 2010).

COOT: For density modification and visualization during structure refinement (Emsley and Cowtan, 2004).

Diffraction Anisotropy Server (http://services.mbi.ucla.edu/anisoscale/): To check for the presence of anisotropy in crystallization data.

2.1.8.5. Modelling of tertiary structure

SWISS-MODEL (http://swissmodel.expasy.org): To model unknown 3D structure from protein sequences (Arnold et al., 2006).

2.1.8.6. SAXS analysis

ATSAS software package: To analyze SAXS data and calculate SAXS models (Petoukhov et al., 2012).

2.1.8.7. EM analysis

IMOD: To preprocess micrographs (bio3d.colorado.edu/imod/)

BSoft: To preprocess micrographs (Heymann and Belnap, 2007).

TiltPicker : To pick particle from tilt pairs of micrographs (Voss et al., 2009).

IMAGIC: To perform 2D classification (van Heel et al., 1996).

XMIPP: software package: To perform CTF correction on preprocessed micrographs and to do 3D reconstruction (Sorzano et al., 2004).

SPIDER: To perform reprojection from 3D models (Shaikh et al., 2008).

2.1.8.8. Protein structure visualization

PyMOL: To visualize protein structures. (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.)

UCSF Chimera: To visualize EM models (Pettersen et al., 2004).

2.2. Methods

2.2.1. Preparation of buffers

All buffers and solutions were prepared using Milli-Q water (Millipore systems) and filtered with a 0.22 µm filter. HCl or NaOH were used for adjusting pH, if not otherwise stated. Buffers for DNA polymerases and restriction enzymes were from New England Biolabs (5x or 10x stock solutions).

2.2.2. General nucleic acid biochemistry

2.2.2.1. Isolation of plasmid DNA

Plasmid DNA was purified from 5-20 mL bacterial cultures using the QIAprep Spin Miniprep Kit following the manufacturer's manual. Typically 40-50 µL of plasmid DNA was eluted and stored at -20°C. For culture volumes >10 mL, double volumes of the buffers P1, P2 and N3 were used.

2.2.2.2. Concentration determination of nucleic acids

The nucleic acid concentration was determined in aqueous solutions by measuring the absorbance at 260 nm against a reference solution with a Nanodrop 1000 spectrophotomer. Concentrations were calculated from absorbance according to the following equation:

$$1\ OD_{260} = 50\ \mu g/ml \text{ doubled stranded DNA}$$

2.2.2.3. Polymerase chain reaction (PCR)

PCR was used to amplify genes of interest and also to amplify plasmid backbones (SLIC procedure, Chapter 2.2.2.9). In general, to prepare the PCR reaction 30-100 ng of template DNA was mixed with 10 µL of 5x HF or 5x GC-Phusion Buffer, 1.5 µL of forward primer and reverse primer each at 10 µM concentration, 1 µL of 10 mM dNTP Mix, up to 3 µL of 100% DMSO, up to 12 µL of 5M Betain, 1 U of Phusion® High-Fidelity DNA Polymerase and topped up to 50 µL total volume with milli-Q H_2O.

A thermocycler was used to regulate the temperatures required for the reaction to occur. The template was initially denatured at 98°C for 60 sec, followed by 30x cycles of a denaturing step at 98°C for 30 sec, an annealing step at a temperature 2-5°C lower than the melting temperature of the primers used in the reaction for 30 sec and an elongation step at 72°C for 45 sec per kb of gene to be amplified. Following this there was a final elongation step at 72°C for 5-10 min. The PCR was stored at 4°C (short-term) or -20°C (long-term) until further use.

2.2.2.4. PCR purification

PCR purification was performed using PCR purification kit following the manufacturer's instructions.

2.2.2.5. Agarose gel electrophoresis

Agarose gel electrophoresis was used to analyze DNA molecules. Agarose gels were prepared by adding 0.5-2 % agarose in 1x TBE (178 mM Tris, 178 mM Boric acid, 4 mM EDTA, pH 8.0) depending on size of the DNA molecules to be analyzed. For visualization of the DNA bands by UV light, Ethidium bromide was added in a 1:20000 [v/v] ratio before gel solidification. To load on the agarose gels, DNA samples were mixed in a 5:1 [v/v] ratio with loading dye (6x BX-DNA loading dye: 30 % [v/v] glycerol, 0.125 % [w/v] bromophenol blue, 0.125 % [w/v] xylene cyanol FF) and separated at 80-120 V for 30-120 min depending on percentage of agarose in the gel and size of DNA molecules.

2.2.2.6. Restriction digestion of DNA

Restriction digestion was used to generate compatible sticky ends in PCR products and vectors for subsequent ligation or to validate recombinant plasmids by restriction mapping. Restriction digestions were carried out according to the manufacturer's manuals. In general, 1 µg of plasmid DNA were digested with 1 U of enzyme for 1.5 hr at the 37°C.

2.2.2.7. DNA extraction from agarose gels

DNA bands were excised from agarose gels using sterile scalpel blades on an UV table at 365 nm wavelength. The DNA was extracted from the excised agarose gel slices using a QIAquick Gel Extraction Kit following the manufacturer's instructions.

2.2.2.8. DNA ligation

Ligation was performed to ligate compatible sticky ends of vector and insert using T4 DNA Ligase. Typically ligation reactions were prepared by mixing ~200 ng of vector with insert in a 1:3 molar ratio, supplemented with 1 µL 10x T4 DNA Ligase Buffer and 10 U of T4 DNA Ligase in a 10 µL reaction and incubated at room temperature for 1-2 hr. 5-10 µL of ligation reaction was used for transformation into appropriate chemically competent cells.

2.2.2.9. Sequence and ligation independent cloning (SLIC)

SLIC was performed as described by (Li and Elledge, 2007). In brief, 1 µg of each linear insert and linearized vector (prepared by PCR amplification) were separately treated with 0.5 U of T4 DNA Polymerase in a 20 µL reaction supplemented with 2 µL 10x NEBuffer2.1 at 30°C for 30-50 min, depending on the length of overhangs. The reaction was stopped by adding 2 µL of 10 mM dCTP and kept on ice. ~200 ng of this T4 DNA Polymerase treated vector was mixed with T4 DNA Polymerase treated insert in a 1:2 molar ratio in a 10 µL reaction supplemented with 1 µL of 10x T4 DNA Ligase buffer and incubated at 37°C for 60 min. 5-10 µL of ligation reaction was used for transformation into appropriate chemically competent cells.

2.2.2.10. Cre-LoxP recombination of Acceptor and Donor plasmids

Cre-LoxP recombination was used to fuse donor and acceptor plasmids. Typically, a 10 µL Cre-LoxP reaction was prepared by mixing 0.5-1 µg of Acceptor with a 1:1.1 molar ratio of Donor(s), along with 1 µL 10x Cre Buffer (500 mM Tris/HCl pH 7.5, 330 mM NaCl, 100 mM $MgCl_2$) and 1 µL Cre recombinase. The reaction was incubated at 37°C for 1 hr,

followed by transformation into appropriate chemically competent cells using 5-10 µL of Cre-LoxP reaction mix.

2.2.3. Cells and cell culture

2.2.3.1. Cultivation of Escherichia coli (*E.coli*)

For growing *E. coil* cells LB medium or LB agar plates supplemented with appropriate antibiotics were used. Cell growth was checked by measuring optical densities (ODs) of cultures using plastic cuvettes with 1 cm path-length in a Biowave 3000 CO8000 Cell Density Meter at 600 nm wavelength against LB medium as a reference.

2.2.3.2. Transformation of bacterial cells

Chemically competent cells were transformed using heat shock, where 5-10 µL of ligation or Cre-LoxP recombination reaction or purified plasmid (~100-200ng) was added to 50-100 µL of competent cells and incubated on ice for 30 min. Then the heat shock was given to the cells at 42°C for 45-60 sec followed by incubation on ice for 2 min. Then, 400 µL of LB medium was added to the cells and incubated at 37°C in a shaker for 1-2 hr or overnight followed by plating on LB agar plates supplemented with appropriate antibiotics to select positive tranformants. Antibiotic concentrations used were 30 µg/mL for Kan, 34 µg/mL for Camp, 15 µg/mL for Gent, 100 µg/mL for Amp and 12.5 µg/mL for Tet.

While electro competent cells were transformed using electroporation by an electric pulse, where plasmid or ligation/Cre-LoxP recombination reactions (it was desalted to remove any salts, which might create arcing during electro transformation resulting in failure) were mixed with electro competent cells similar to chemical transformation. This mix was transferred to an electroporator cuvette and kept on ice. Then electric pulse was given to the cells using Bio-Rad *E. coli* Pulser at 1.8~2.0 kV followed by immediate addition of 400 µL of LB medium. Next, these cells were incubated and plated in a similar way as described for chemical transformation.

2.2.3.3. Cultivation and maintenance of insect cells

Sf21 insect cells were grown and maintained in insect cell medium. Cells were always maintained between cell counts of $0.4x10^6$ – $3.0x10^6$. Cells were counted using a neubauer cell counter.

2.2.4. General protein biochemistry

2.2.4.1. Denaturing polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-polyacrylamide gel electrophoresis (SDS-PAGE) with 11-18 % Tris-Glycine or Tris-Tricine gels (Schagger, 2006) were used to analyze proteins, depending on protein size and required resolution range. Gels were casted manually containing a separating gel layer with higher percentage acrylamide (11-18 %) covered on top by a stacking gel layer with lower acrylamide percentage (6 % or 4 %, respectively).

In Tris-Glycine SDS-PAGE, the stacking layer composition was 125 mM Tris/HCl (pH 6.8), 0.1 % [w/v] SDS, 6 % acrylamide / bis-acrylamide (37.5:1), which was polymerized by 0.4 % [v/v] APS / 0.05 % [v/v] TEMED. The separating layer composition was 375 mM Tris/HCl (pH 8.8), 0.1 % SDS, 10-16 % acrylamide / bis-acrylamide (37.5:1), polymerized by 0.4 % [v/v] APS / 0.05 % [v/v] TEMED.

In Tris-Tricine SDS-PAGE, the stacking layer composition was 1 M Tris/HCl (pH 8.45), 0.1 % [w/v] SDS, 4 % acrylamide / bis-acrylamide (29:1) which was polymerized by 0.5 % [v/v] APS / 0.08 % [v/v] TEMED. The separating layer composition was 1 M Tris/HCl (pH 8.45), 0.1 % SDS, 10-16 % acrylamide / bis-acrylamide (29:1), polymerized by 0.5 % [v/v] APS / 0.08 % [v/v] TEMED.

For loading on gel, protein samples were prepared by mixing with 4x protein loading dye (200 mM Tris/HCl pH 6.8, 8 % [w/v] SDS, 40 % glycerol, 4 % [v/v] β-ME, 50 mM EDTA and 0.08 % [w/v] bromophenol blue) in 1:3 dye to protein [v/v] ratio and denatured by heating at 95°C for 2 min. SDA-PAGE running buffer for Tris-Glycine SDS-PAGE was 25 mM Tris/HCl pH 8.2, 192 mM glycine, 0.1 % [w/v] SDS, while 1x Anode buffer

(100 mM Tris/HCl pH 8.9) and 1x Cathode buffer (100 mM Tris/HCl, 100 mM Tricine, 0.1 % SDS) were used to run Tris-Tricine SDS-PAGE. Coomassie staining (40 % [v/v] Ethanol, 10 % [v/v] acetic acid and 0.2 % [w/v] Coomassie Brilliant Blue G-250) was used to visualize protein bands. De-staining solution containing 5 % [v/v] Ethanol and 7.5 % [v/v] acetic acid was used to remove background staining.

2.2.4.2. Protein expression in insect cells with the MultiBac system

Most of the protein complexes were expressed using MultiBac baculovirus/insect cell expression system in *Sf*21 insect cells (Berger et al., 2004; Bieniossek et al., 2008; Fitzgerald et al., 2006). Briefly, bacmid was isolated from DH10-EMBacY cells and was used to transfect *Sf*21 insect cell in 6-well plates in a total culture volume of 3 mL per well with a total cell count of $0.5x10^6$ to produce first generation V_0 virus. This virus was used to infect 25-50 mL *Sf*21 cell cultures in shaker flasks to produce higher titer second generation V_1 virus. For large-scale protein expression V_1 virus was used to infect 400 mL *Sf*21 cell cultures in 2 L shaker flasks. Yellow fluorescent protein (YFP) was used as a marker of protein expression and when it reached a plateau, cells were harvested. YFP typically reached plateau at 72-96 hours after the day of proliferation arrest (DPA). See Publication 2, Publication 3, Publication 4 and Publication 5 for detailed protocols.

2.2.4.3. Protein production in bacterial cells

Some of the proteins were expressed using *E.coli* strains BL21 (DE3) star and Rosetta (DE3). Cells transformed with plasmid of interest were grown at 37°C to an OD of 0.6-0.8 at 600 nm from a starting 0.05-0.08 OD. Then the protein expression via T7 promoter was induced with 0.5-1.0 mM IPTG for either 2-3 hours at 30°C or for overnight at 18°C. ACEMBLE system was used to coexpress protein complexes in bacterial cells (Bieniossek et al., 2009).

2.2.4.4. Labelled protein production in bacterial cells for NMR

Proteins labelled with ^{15}N and/or ^{13}C for NMR studies were expressed using *E. coli* strain BL21 star (DE3) in M9 medium supplemented with ^{15}N labelled NH_4Cl and/or ^{13}C labelled glucose. M9 medium was prepared mixing 3g KH_2PO_4, 8.5g $Na_2HPO_4.2H_2O$, 0.5g NaCl in 1L Milli-Q water and autoclaved. Before using 1g NH_4Cl, 2g glucose, 10 mL MEM vitamins mix 100X, 2 mL 1M MgSO4, 100µl 1M CaCl2 and appropriate antibiotics were added to prepare complete M9 medium. NH_4Cl and glucose were replaced with corresponding labelled chemicals as per requirement. First the cells were grown at 37°C to an OD of 0.6-0.8 at 600 nm in LB medium, then LB medium was removed and cells were washed and resuspended in complete M9 medium. Cells from 4L LB medium were resuspended in 1L complete M9 medium. Then cells were again grown for 1-1.5 h at 37°C and then induced with 1mM IPTG for overnight at 18°C.

2.2.4.5. Concentration determination of proteins

Protein concentrations were determined by measuring absorbance at 280 nm against a reference buffer with a Nanodrop 1000 spectrophotomer. From this absorbance, protein concentrations were calculated using appropriate extinction coefficients calculated from the protein sequence with webserver ProtParam.

2.2.4.6. Western blot

Western blots were mainly used to confirm the expression of proteins and protein complexes. All constructs, used in this study, contained His-tag. Standard procedures were used, where proteins were first transferred to PVDF membrane using a semidry apparatus and then after blocking, were detected using specific antibodies for His-tag.

2.2.4.7. Protein purification

Protein purifications involved the use of different kinds of chromatography. IMAC with Talon resin or Ni^{2+}-NTA resin was performed using AKTA prime with XK16 columns or

Bio-Rad gravity flow columns. AKTA Purifier systems were used for IEX with MonoQ 5/50GL or MonoS 5/50GL or HiTrap SPFF columns and SEC with SuperdexS200 10/300, SuperdexS200 16/60 or Superose6 10/300 columns. Flow rates were typically 1-2 mL/min MonoQ 5/50GL, MonoS 5/50GL, HiTrapQ and HiTrap Heparin columns, 1 mL/min for HiTrap SPFF, SuperdexS75 16/60 and SuperdexS200 16/60 columns, 0.5 mL/min for SuperdexS75 10/300, SuperdexS200 10/300 and Superose6 10/300 columns. All steps for all protein purifications were carried out at 4°C. AKTA Micro system, at 50 µL/min flow rate, was used for SEC with SuperdexS200 PC3.2/30 column.

2.2.4.8. Protein crystallization

Protein samples were screened for initial crystallization conditions with different commercial crystallization screens, using an automated crystallization plate imaging system at the HTX laboratory, EMBL Grenoble. The crystallization drops were set up in 96-well sitting drop plates by mixing 100-150 nL protein sample with reservoir solution in 1:1 ratio, which were equilibrated against 70 µL reservoir solution.

Conditions, which produced crystals in initial screening, were optimized for optimal crystal growth by changing different parameters- pH, salts, salt and precipitant and protein concentrations, protein and reservoir solution etc. For optimization, crystallization drops were set up in 24-well hanging or sitting drop crystallization plates by mixing 1-2 µL of protein sample with reservoir solution in 1:1 ratio, and then equilibrated against 500 µL of reservoir solution. Growth and formation of crystals was further optimized using crystal seeding techniques- streak-seeding, micro-seeding. All crystals were grown at 4°C.

2.2.5. Methods for characterization of TAF1/TAF7 complex

2.2.5.1. Purification of $TAF1^{578-1210}/TAF7^{1-193}$ and $TAF1^{578-1106}/TAF7^{1-193}$

Both TAF1/TAF7 constructs were coexpressed in *Sf*21 insect cells using corresponding plasmid. 2L of cell culture was harvested at ~DPA+72 by centrifuging at 1100g using

JA8.100 rotor for 10 min. Cells were lysed in Lysis buffer/Talon A buffer (25mM Tris pH 8.0, 300mM NaCl, 5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) (10mL/gm of cell pellet) by freeze-thaw (freezing in liquid nitrogen and thawing on ice for two times), followed by centrifugation at 19000 rpm using JA20.0 rotor for 1 hr. The supernatant was loaded on 5mL talon resin, pre-equilibrated in 10 CV of Talon A buffer (pre-packed in HK16 column) using an ÄKTA prime at flow rate of 2mL/min. Then the resin was washed with 10 CV Talon A buffer, followed by 10 column volume Talon HS buffer (25mM Tris pH 8.0, 1M NaCl, 5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet), then again 10 CV Talon A buffer. Afterwards, complex was eluted using a continuous gradient of Talon buffer B (25mM Tris pH 8.0, 300mM NaCl, 200mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-100% in 60mL, followed by 20mL gradient delay. 0.5mg TEV protease was added to the eluted sample, followed by overnight (O/N) dialysis against Dialysis buffer (25mM Tris pH 8.0, 300 mM NaCl, and 2mM β-ME), using dialysis bag of 6-8 kDa MW cutoff. This dialyzed sample was incubated with 1mL of pre-equilibrated Talon resin for 45 min and flow-through (FT) was collected, which contained TAF1/TAF7 complexes devoid of His-tag. This FT was concentrated to 0.5 mL using amicon spin concentrators with a 10 kDa MW cutoff membrane and injected on SuperdexS200 10/300, pre-equilibrated with SEC buffer at 0.5 mL/min in SEC buffer (25mM Tris pH 8.0, 300 mM NaCl, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) with a fraction size of 400 µL. SEC buffer for $TAF1^{578\text{-}1106}/TAF7^{1\text{-}193}$ contained 150mM NaCl (instead of 300mM). Fractions containing the corresponding complex were pooled and concentrated using amicon spin concentrators with a 10 kDa MW cutoff membrane. Concentrated protein complexes were stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.5.2. Crystallization, X-ray data collection and structure determination of TAF1$^{578-1106}$/TAF7$^{1-193}$ complex

Purified TAF1$^{578-1106}$/TAF7$^{1-193}$ complex was screened for initial crystallization condition at HTX laboratory (EMBL Grenoble). The main precipitant in those conditions was PEG3350 and these conditions were further optimized for better crystal growth. Complex was crystallized in final optimized buffer containing 0.2 M di-Sodium Malonate pH 7.0, 15% [w/v] PEG 3350, using the hanging drop vapor diffusion technique by mixing 1 µL protein solution at 6.8–7.4 mg/mL with 1 µL of reservoir solution and equilibrated against 500 µL reservoir solution at 4°C. First crystals of rectangle shape appeared after three days and reached their maximal size within 15 days.

For data collection at cryogenic temperatures, crystals were transferred into 2 µL drops containing 0.2 M di-Sodium Malonate pH 7.0, 15% [w/v] PEG 3350 and 25% Ethylene Glycol (or 20% Glycerol) before they were flash frozen in liquid nitrogen.

Native datasets were collected mainly at beamline ID29 (ESRF, Grenoble, France (de Sanctis et al., 2012)) at 100 K and also at ID23-1 (ESRF, Grenoble, France (Nurizzo et al., 2006)) at 100 K or ID23-2 (ESRF, Grenoble, France (Flot et al., 2010)) at 100 K. Data collection strategies were calculated from two initial diffraction images using EDNA (Leal et al., 2011) as implemented in MXCuBE (Gabadinho et al., 2010) according to observed space group, cell parameters and X-ray dose to minimize radiation damage.

All diffraction data were processed with the XDS software package. Data were checked for anisotropy using diffraction anisotropy webserver. The structure was solved by molecular replacement (MR) using a poly Alanine model of the Yeast TAF1/TAF7 (PDB-ID 4OY2 (Bhattacharya et al., 2014)). The initial MR solution was checked and manually modified in COOT by deleting or moving parts of the structure not occupying 2Fo-Fc density >1.5σ. The model was manually adjusted in repetitive rounds of refinement in PHENIX and manual model building in COOT. TLS refinement (Painter and Merritt, 2006) was used in the final rounds of refinement, with a total number of 17 individual TLS groups for all

protein chains as determined by PHENIX. The final structure of the complex was validated by MOLPROBITY (Chen et al., 2010; Davis et al., 2007) and the PDB Validation Server.

2.2.5.3. Purification of TAF1$^{1157-1207}$/TAF7$^{126-202}$ and TAF1$^{1157-1207}$/TAF7$^{153-190}$ complexes with^{15} N and/or ^{13}C labeling for NMR

^{15}N or ^{15}N-^{13}C labelled or unlabeled TAF1 and TAF7 were expressed separately in *E.coli* BL21 (DE3) star cells using corresponding plasmids, induced with 1mM IPTG for overnight at 18°C and 1L of cell culture was harvested by centrifuging at 6000 rpm using JA8.100 rotor for 20 min. TAF1 and TAF7 were purified separately. Cell were lysed in Lysis buffer (25mM Tris pH 7.5, 300mM NaCl, 5mM imidazole, 1mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) (10mL/gm of cell pellet) supplemented with 0.1 mg/mL lysozyme, using sonication (amplitude 80V, pulse ON 5sec, pulse OFF 10 sec, total process time 180 sec, on ice) followed by centrifugation at 21500 rpm using JA25.5 rotor for 1 hr. The supernatant was loaded on pre-packed Talon cartridge (1mL resin), pre-equilibrated in 20 CV of Talon A buffer (25mM Tris pH 7.5, 300mM NaCl, 5mM imidazole, 1mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) using ÄKTA Prime at flow rate of 2mL/min. Then the resin was washed with 20 CV Talon A buffer, followed by 10 CV Talon HS buffer (25mM Tris pH 7.5, 1M NaCl, 7.5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet), then again 20 CV Talon A buffer. Afterwards, proteins were eluted with a continuous gradient of Talon buffer B (25mM Tris pH 8.0, 1M NaCl, 300mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-100% in 40 CV, followed by 10 CV gradient delay. Elutions containing the proteins were pooled and dialysed O/N against SEC buffer (25mM HEPES pH 7.0, 300 mM NaCl, 1mM EDTA, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet), using dialysis cassettes of 3 kDa (TAF7) or 10kDa (TAF1) MW cutoff. This dialyzed sample was concentrated using amicon spin concentrators with a 3 kDa (TAF7) or 10kDa (TAF1) MW cutoff membrane to appropriate volume and injected on SuperdexS75 10/300 (or16/60)

[TAF7] or SuperdexS200 10/300 (or 16/60) [TAF1], pre-equilibrated with SEC buffer with appropriate flowrate to collect elution fractions of 0.4mL or 1mL. Fractions containing the proteins were pooled and concentrated using amicon spin concentrators with a 3 kDa (TAF7) or 10kDa (TAF1) MW cutoff membrane. Concentrated proteins were stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

Afterwards purified TAF1 and TAF7 were mixed in equimolar ratio to form complex and then incubated with TEV protease at protease/protein ratio of 1:20 [w/w] at 4°C for O/N. Cleaved sample was loaded on MonoS 1mL (TAF1$^{1157-1207}$/TAF7$^{126-202}$) or HiTrap SPFF 1mL (TAF1$^{1157-1207}$/TAF7$^{153-190}$) column pre-equilibrated with IEX buffer A (25mM HEPES pH 7.0, 300 mM NaCl, 1mM EDTA, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet), followed by 15CV wash with IEX buffer A and then elution in a continuous gradient of IEX buffer B (25mM HEPES pH 7.0, 1M NaCl, 1mM EDTA, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-100% in 60 CV. Elutions containing the protein complex were buffer exchanged with SEC buffer (25mM HEPES pH 7.0, 300 mM NaCl, 1mM EDTA, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) as well as concentrated using amicon spin concentrators with a 3 kDa MW cutoff membrane to 0.5mL. This sample was injected on SuperdexS75 10/300, pre-equilibrated with SEC buffer at 0.5mL/min flowrate to collect elution fractions of 0.4mL. Fractions containing the proteins were pooled and concentrated using amicon spin concentrators with a 3 kDa MW cutoff membrane. Concentrated protein complexes were stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.5.4. NMR analysis of TAF1$^{1157-1207}$/TAF7$^{126-202}$ and TAF1$^{1157-1207}$/TAF7$^{153-190}$

Purified ^{15}N or ^{15}N- ^{13}C labelled TAF1/TAF7 complexes (concentration ~190μM-560μM) were provided for NMR analyses, which were carried out in collaboration with Martin Blackledge and Malene Ringkjøbing Jensen (IBS, Grenoble).

All NMR experiments were carried out at a temperature of 25°C (or 10°C) on Bruker spectrometers operating at ^{1}H frequencies of 700 or 950 MHz. The backbone assignment of the resonances of the TAF1/TAF7 complex was obtained employing two different samples of differential labelling (^{15}N, ^{13}C TAF1 in complex with unlabelled TAF7 and ^{15}N, ^{13}C TAF7 in complex with unlabelled TAF1). Both samples contained 0.5 mM TAF1/TAF7 complex in SEC buffer (25mM HEPES pH 7.0, 300 mM NaCl, 1mM EDTA, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet). The backbone assignment was obtained using a series of BEST-type triple resonance experiments: HNCO, intra-residue HN(CA)CO, HN(CO)CA, intra-residue HNCA, HNCOCACB and intra-residue HNCACB (Lescop et al., 2007). All spectra were processed using NMRPipe (Delaglio et al., 1995) and analyzed in Sparky (T.D. Goddard, D.G. Kneller, SPARKY 3, University of California, San Francisco). Automatic assignment of spin systems was done using MARS (Jung and Zweckstetter, 2004) followed by manual verification.

Assignment of side chains and collection of distance restraints for structure determination were obtained by manual inspection of the following three-dimensional spectra: ^{1}H, ^{15}N-TOCSY-HSQC (mixing time of 60 ms), ^{1}H, ^{15}N-NOESY-HSQC (mixing time of 100 ms), ^{1}H, ^{13}C-NOESY-HSQC (mixing time of 100 ms) and ^{1}H-^{13}C HCCH-TOCSY. A 0.5 mM sample of uniformly ^{15}N, ^{13}C labelled TAF1/TAF7 complex was used for these experiments. A total of 588 NOEs were manually assigned and divided into three classes (strong, medium and weak) according to their intensities and converted into distance restraints with upper limit distances of 3.7 Å (strong), 4.3 Å (medium) and 6.0 Å (weak). Standard pseudo-atom corrections were used for unambiguously assigned protons.

An ensemble of structures was calculated using CNS (Brunger et al., 1998) on the basis of the collected distance restraints including water refinement. Dihedral angle restraints were obtained using TALOS+ (Shen et al., 2009) employing an error of twice the TALOS error. Hydrogen bond restraints were included in regions of regular secondary structure after

initial structure calculations. Final structures were validated using PROCHECK (Laskowski et al., 1996).

2.2.5.5. SAXS data collection and analysis of TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ complexes

Small-angle X-ray scattering experiments were carried out at the ESRF BioSAXS beamline BM29 (Pernot et al., 2013), Grenoble, France with the help from Adam Round, EMBL Grenoble, using program BSXcuBE. 30µl for each of purified TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ at different concentrations (see Appendix and Supplements, Table 13) along with sample buffer (25mM Tris pH 8.0, 300 mM NaCl, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) were exposed to X-rays and scattering data collected using the robotic sample handling available at the beamline. SEC analysis was considered as a check for monodispersity. Ten individual frames were collected for every exposure, each 2 seconds in duration using the Pilatus 1M detector (Dectris). Data processing was done with different programs of ATSAS software package. Individual frames were processed automatically and independently within the EDNA framework, yielding individual radially averaged curves of normalized intensity versus scattering angle $S=4\pi Sin\theta/\lambda$. Additional data reduction within EDNA utilizes the automatic data processing tools of ATSAS software package, to combine timeframes, excluding any data points affected by aggregation induced by radiation damage, yielding the average scattering curve for each exposure series. Matched buffer measurements taken before and after every sample were averaged and used for background subtraction. Merging of separate concentrations and further analysis steps were performed manually using the PRIMUS (Konarev et al., 2003). PRIMUS was used to calculate the forward scattering (I(0)) and radius of gyration (Rg) from the Guinier approximation, to compute the hydrated particle volume using the Porod invariant and to determine the maximum particle size (Dmax) from the pair distribution function computed by GNOM. 40 *ab initio* models were calculated for each sample, using

DAMMIF (Franke and Svergun, 2009), and then aligned, compared and averaged (which showed minimal variation) using DAMAVER (Volkov and Svergun, 2003). Rigid body models were generated for both constructs using crystal structure of TAF1$^{578\text{-}1210}$/TAF7$^{1\text{-}193}$ and/or NMR structure of TAF1$^{1157\text{-}1207}$/TAF7$^{153\text{-}190}$ to also model missing loops and residues using CORAL (Petoukhov et al., 2012). The most representative model for TAF1$^{578\text{-}1210}$/TAF7$^{1\text{-}193}$ and TAF1$^{578\text{-}1106}$/TAF7$^{1\text{-}193}$ selected by DAMAVER were compared to corresponding rigid body models, with overlays of the resulting models created with PYMOL. The fits to the experimental data of the models and the theoretical scattering of the calculated structures were generated with CRYSOL (Svergun et al., 1995).

2.2.5.6. Peptide Array

Pepscan libraries of the TAF7$^{156\text{-}190}$ were immobilized on cellulose membranes via double b-alanine anchors and assembled using the SPOT technology (AG Molekulare Bibliotheken, Charite Universitatsmedizin Berlin, Germany). Different 35-mer peptides of TAF7$^{156\text{-}190}$ – one wild type and 35 others by mutating each AA to Alanine (to Serine wherever Alanine was present in WT sequence, one at a time), were synthesized by Fmoc (9-fluorenylmethoxycarbonyl) chemistry. Low-density hexa-Histidine peptides were used as controls. Pepscan membranes were blocked in blocking buffer (50mM Tris-HCl, pH 7.6, 300mM NaCl, 20% (w/v) sucrose, 3% (w/v) bovine serum albumin) for 1 h at 4 C, washed with TBS (50mM Tris-HCl, pH 7.6, 300mM NaCl) and incubated for 2 hr with C-terminally His-tagged MBP_TAF1$^{1157\text{-}1207}$ or His_MBP as a control at 30 µg/mL each in blocking buffer. Membranes were incubated with mouse anti-His monoclonal primary antibody at a dilution of 1:3000, followed by peroxidase-conjugated anti-mouse secondary antibody at a dilution of 1:10000 in blocking buffer. Membranes were washed three times with TBS between each incubation step. Luminol solution (Pierce) was added and luminescence detected on a Bio-Rad ChemidocMP. Images were analyzed using the Dot Blot Analyzer tool in ImageJ. Data was normalized against the background signal from His_MBP control.

2.2.6. Methods for characterization of TAF11/TAF13/TBP complex

2.2.6.1. Purification of TAF11/TAF13 complex

The TAF11/TAF13 complex was co-expressed in *Sf*21 insect cells using plasmid pFB_6xHis_TEV_TAF11_TAF13. 2L of cell culture was harvested at ~DPA+72 by centrifuging at 1100g in a JA8.100 rotor for 10 min. Cells were lysed in Lysis buffer/Talon A buffer (25mM Tris pH 8.0, 150mM NaCl, 5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) (8mL/gm of cell pellet) by freeze-thaw, followed by centrifugation at 40000g in Ti70 rotor for 1.5 hr. The supernatant was loaded on 4mL talon resin, pre-equilibrated in 10 CV of Talon A buffer (pre-packed in HK16 column) using an ÄKTA prime at flow rate of 2mL/min. Then the resin was washed with 10 CV Talon A buffer, followed by 10 CV Talon HS buffer (25mM Tris pH 8.0, 1M NaCl, 5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet), then again 10 CV Talon A buffer. Afterwards, TAF11/TAF13 was eluted using a continuous gradient of Talon buffer B (25mM Tris pH 8.0, 150mM NaCl, 200mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-100% in 100mL, followed by 20mL gradient delay. Fractions containing the TAF11/TAF13 complex were pooled and dialyzed O/N against HiTrapQ buffer A (50mM Tris pH 8.0, 150mM NaCl, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet), using dialysis bag of 6-8 kDa MW cutoff. Dialyzed sample was loaded on HiTrapQ column (5mL), pre-equilibrated with 5 CV of HiTrapQ buffer A at a flow rate of 2mL/min. After binding, column was washed with 5 CV HiTrapQ buffer A and then TAF11/TAF13 was eluted using a continuous gradient of Talon buffer B (50mM Tris pH 8.0, 1M NaCl, 5mM β-ME, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-50% in 80mL, followed by a step gradient of 50 to 100% for 50mL. Fractions containing the TAF11/TAF13 complex were pooled and dialyzed O/N against SEC buffer (25mM Tris pH 7.5, 150 mM NaCl, 1mM EDTA, 1mM DTT, Leupeptin, Pepstatin and complete protease inhibitor tablet), using dialysis bag of 6-8 kDa MW cutoff. This

dialyzed sample was concentrated to 0.5mL using amicon spin concentrators with a 3 kDa MW cutoff membrane and injected on SuperdexS75 10/300, pre-equilibrated with SEC buffer at 0.5 mL/min with a fraction size of 400 µL. Fractions containing the TAF11/TAF13 complex were pooled and concentrated using amicon spin concentrators with a 3 kDa MW cutoff membrane. Concentrated protein complex was stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.6.2. Purification of TBP core

Conserved core of TBP was expressed in *E.coli* Rosetta (DE3) using plasmid pET28(a)_6xHis_TBPcore, induced with 1mM IPTG for overnight at 18°C. 6L of cell culture was harvested by centrifuging at 6000 rpm using a JA8.100 rotor for 20 min. Cells were lysed in Lysis buffer/Talon A buffer (25mM Tris pH 8.0, 1M NaCl, 10mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) (6mL/gm of cell pellet) by using French-press at 1100 kPSI for 3 times, followed by centrifugation at 40000g using Ti70 rotor for 1.5 hr. The supernatant was loaded on 6mL talon resin, pre-equilibrated in 10 CV of Talon A buffer (pre-packed in HK16 column) using an ÄKTA prime at flow rate of 2mL/min. Then the resin was washed with 15 CV Talon A buffer, followed by elution using a continuous gradient of Talon buffer B (25mM Tris pH 8.0, 1M NaCl, 600mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) from 0-100% in 100mL, followed by 20mL gradient delay. Fractions containing the TBPc were pooled and dialysed O/N against SEC buffer (25mM Tris pH 8.0, 300 mM NaCl, 1mM EDTA, 1mM DTT, Leupeptin, Pepstatin and complete protease inhibitor tablet), using dialysis bag of 6-8 kDa MW cutoff. The dialyzed sample was concentrated to 1.0 mL using amicon spin concentrators with a 3 kDa MW cutoff membrane and injected on SuperdexS75 16/60, pre-equilibrated with SEC buffer at 1.0 mL/min with a fraction size of 1 mL. Fractions containing the TBPc were pooled and concentrated using amicon spin concentrators with a 3

kDa MW cutoff membrane. Concentrated protein was stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.6.3. Purification of TBP-FL

TBP-FL was expressed in *E.coli* BL21 (DE3) star using plasmid pET28(a)_6xHis_TEV_TBP-FL, induced with 0.5mM IPTG for 2 hours at 30°C. 6L of cell culture was harvested by centrifuging at 6000 rpm using JA8.100 rotor for 20 min. Cells were lysed in Lysis buffer/Talon A buffer (25mM Tris pH 8.0, 1M NaCl, 5mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet) (6mL/gm of cell pellet) by using French-press at 1100 kPSI for 3 times, followed by centrifugation at 19000rpm using JA25.5 rotor for 45 min. The supernatant was incubated for 1hr with 4mL talon resin, pre-equilibrated in 10 CV Talon A buffer, in a Bio-Rad gravity flow column. Then the resin was washed with 15 CV Talon A buffer, followed by elution with 5 CV Talon B buffer (25mM Tris pH 8.0, 1M NaCl, 300mM imidazole, Leupeptin, Pepstatin and complete protease inhibitor tablet). ~0.5mg TEV protease was added to the eluted sample, followed by O/N dialysis against Dialysis buffer (25mM Tris pH 8.0, 300 mM NaCl, 5mM β-ME), using dialysis bag of 6-8 kDa MW cutoff. This dialyzed sample was loaded on 1mL of pre-equilibrated Ni^{2+}-NTA resin and FT was collected, which contained TBP-FL devoid of his-tag. The FT was concentrated to 0.5 mL using amicon spin concentrators with a 30 kDa MW cutoff membrane and injected on SuperdexS200 10/300, pre-equilibrated with SEC buffer (25mM Tris pH 8.0, 300 mM NaCl, 1mM EDTA, 1mM DTT, Leupeptin, Pepstatin and complete protease inhibitor tablet) at flow rate 0.5 mL/min with a fraction size of 400 µL. Fractions containing the TBP-FL were pooled and concentrated using amicon spin concentrators with a 30 kDa MW cutoff membrane. Concentrated protein was stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.6.4. Purification of scTFIIA

scTFIIA was purified using the protocol provided by Eaazhisai Kandiah, Berger group. Protein was expressed in *E.coli* BL21 (DE3) star using plasmid pET28(a)_6xHis_scTFIIA, induced with 1mM IPTG for overnight at 18°C. 3L of cell culture was harvested by centrifuging at 6000 rpm using JA8.100 rotor for 20 min. Cells were lysed in Lysis buffer/Talon A buffer (20mM Tris pH 7.4, 150mM NaCl, complete protease inhibitor tablet) (10mL/gm of cell pellet) by using French-press at 1100 kPSI for 3 times, followed by centrifugation at 20000 rpm using JA20.0 rotor for 45 minutes. The supernatant was loaded on 3mL talon resin, pre-equilibrated in 10 CV of Talon A buffer (pre-packed in HK16 column) using an ÄKTA prime at flow rate of 2mL/min. Then the resin was washed with 10 CV of Talon A buffer, followed by elution using a continuous gradient of Talon buffer B (20mM Tris pH 7.4, 150mM NaCl, 250mM imidazole, complete protease inhibitor tablet) from 0-100% in 60mL, followed by 20mL gradient delay. Fractions containing scTFIIA were pooled and dialyzed O/N against Heparin buffer A (20mM Tris pH 7.4, 150mM NaCl, 1mM DTT, 0.5mM EDTA), using dialysis bag of 6-8 kDa MW cutoff. Dialyzed sample was loaded on HiTrap Heparin column (5mL) pre-equilibrated with 5 CV of Heparin buffer A at a flow rate of 1mL/min. After binding column was washed with 5 CV Heparin buffer A and then scTFIIA was eluted using a continuous gradient of Heparin buffer B (20mM Tris pH 7.4, 1M NaCl, 1mM DTT, 0.5mM EDTA) from 0-100% in 50mL. Fractions containing scTFIIA were pooled and concentrated to 1.0 mL using amicon spin concentrators with a 10 kDa MW cutoff membrane and injected on SuperdexS75 10/300, pre-equilibrated with SEC buffer (20mM Tris pH 7.4, 150mM NaCl, 1mM DTT, 0.5mM EDTA, complete protease inhibitor tablet), at 0.5 mL/min with a fraction size of 0.4 mL. Fractions containing scTFIIA were pooled and concentrated using amicon spin concentrators with a 10 kDa MW cutoff membrane. Concentrated protein was stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.6.5. Preparation of AdMLP dsDNA

Double stranded AdMLP DNA (16 bp DNA oligo duplex containing a TATA box; sequence of coding strand: 5'ctgctataaaaggctg 3') was prepared by mixing equimolar amounts of AdMLP-reverse and –forward ssDNA (BioSpring) in buffer A (10mM Tris pH 8.0, 50mM NaCl, 5mM MgCl2) and incubating the mixture at 95°C for 2 min followed by gradual cooling down to room temperature.

2.2.6.6. EMSA

The EMSA (Electrophoretic mobility shift assay) reactions were made by mixing ds DNA (2uM) with TBPc (4uM) or TBP-FL (4uM), scTFIIA (6uM) and TAF11/TAF13 (2uM-64uM) in EMSA reaction buffer (10mM Tris pH 8.0, 60mM KCl, 10mM MgCl2, 10 % glycerol, 2.5mM DTT) followed by 1.5 hour incubation on ice. To check gel shift, specific non-denaturing polyacrylamide gels, made with only separating layer of 1-1.5mm thickness, were used. The composition of these gels was 5% acrylamide / bis-acrylamide (37.5:1) in 1x EMSA running buffer (25 mM Tris, 190 mM Glycine, 5 mM Mg Acetate, pH 8.0), polymerized by 0.4 % [v/v] APS / 0.05 % [v/v] TEMED. Before loading the samples, gels were pre-run without any samples at 80 V for 60 min at 4C. To load on the gels, samples were mixed in a 5:1 [v/v] ratio with 30 % glycerol, and then separated at 80 V for 60 min at 4C. For visualization of the DNA bands by UV light, gel were stained with 1x TBE (178 mM Tris, 178 mM Boric acid, 4 mM EDTA, pH 8.0) containing ethidium bromide in a 1:10000 [v/v] ratio (15 min incubation), followed by destaining with 1x TBE (15 minutes incubation).

2.2.6.7. Purification/identification of TAF11/TAF13/TBP complex

TAF11/TAF13 was mixed with TBP in 1:1.2 molar ratio in reaction/SEC buffer (25mM Tris pH 8.0, 300 mM NaCl, 1mM EDTA, 1mM DTT, Leupeptin, Pepstatin and complete protease inhibitor tablet) and incubated on ice for 1.5 hr. Then the TAF11/TAF13/TBP complex was separated from TBP excess on SuperdexS200 10/300 column, pre-equilibrated with SEC buffer at 0.5 mL/min with a fraction size of 400 µL. Fractions analyzed using 16% SDS-

PAGE gels. Peak fractions containing the TAF11/TAF13/TBP complex were pooled and concentrated using amicon spin concentrators with a 3 kDa MW cutoff membrane. Concentrated complex was stored at -80°C in small aliquots after flash freezing in liquid nitrogen until further use.

2.2.6.8. Analytical ultra-centrifugation (AUC) of TAF11/TAF13/TBP complex

All AUC experiments were carried out with help from Aurelien Deniaud, EMBL Grenoble. Absorbance at 280nm for purified TAF11/TAF13/TBP complex was measured by spinning it in an An-60Ti rotor (Beckman Coulter) with a Beckman XL-I analytical ultracentrifuge (Beckman Coulter) at 42,000 rpm and 10°C for 16 h. Sapphire-windowed cells with 12-mm optical path length were used for loading the sample. The data were analyzed with the Sedfit program (Schuck, 2000), considering 200 particles with sedimentation coefficients, **S**, between 0.1 and 8 S. A partial specific volume of 0.73, frictional ratio of 1.6 and confidence level of 0.68 were used.

2.2.6.9. Native MS of TAF11/TAF13/TBP complex

All native MS experiments were performed by Ima Obong-Ebong & Prof. Dame Carol Robinson (University of Oxford).

2.2.6.10. Pulldown assay of TAF11/TAF13 with TAF1-TAND for TBP binding

N-terminally MBP tagged TAF1-TAND1 and TAF1-TAND1+2 were expressed in *Sf21* insect cells. 50ml cell pellet of each protein was harvested at ~DPA+72 by centrifuging at 1100g in table top centrifuge for 10 min. Cells were lysed in Lysis buffer/Binding buffer (20mM Tris pH 7.5, 200mM NaCl, 1mM EDTA, 10mM β-ME, Leupeptin, Pepstatin, complete protease inhibitor tablet) by freeze-thaw, followed by centrifugation at 20000 rpm using JA25.5 rotor for 45 min. Supernatant was collected and incubated with 300 µL of Amylose resin for 1hr at 4°C. The FT was collected using gravity flow, followed by extensive washing (20 CV) of resin to remove unbound TAF1-TAND1 (or 1+2). 100 µL of

this resin was then incubated with 1mg of preformed TAF11/TAF13/TBP complex (excess) for 1 hr at 4°C, in a 2mL reaction. Remaining resin with immobilized TAF1-TAND1 (or 1+2) was incubated separately as controls. Subsequently, FT was collected along with washes and finally bound protein(s) were eluted using Elution buffer (20mM Tris pH 7.5, 200mM NaCl, 1mM EDTA, 10mM β-ME, 10mM Maltose, Leupeptin, Pepstatin, complete protease inhibitor tablet). Samples were then analyzed using 16% SDS-PAGE gels.

2.2.6.11. Purification of TAF11/TAF13 deletion mutants and complex formation with TBP

Purifications were done as described in Chapter 2.2.6.1 and complex formations with TBP were checked as described in Chapter 2.2.6.7.

2.2.6.12. SAXS data collection and analysis of TAF11/TAF13/TBP complex

Small-angle X-ray scattering experiments were carried out at the ESRF BioSAXS beamline BM29, Grenoble, France similar to as described in Chapter 2.2.5.5. 30µl of each of purified TAF11/TAF13/TBP, TAF11/TAF13 and TBP at different concentrations (see Appendix and Supplements, Table 14) along with sample buffer (25mM Tris pH 8.0, 300 mM NaCl, 1mM EDTA, 1mM DTT, Leupeptin, Pepstatin and complete protease inhibitor tablet) were used to collect scattering data. Here frames were collected for 2 sec instead of 1 sec as in Chapter 2.2.5.5. Data was processed and analyzed and *ab initio* models were generated as described in Chapter 2.2.5.5 as well. The most representative model for TAF11/TAF13/TBP and TAF11/TAF13 selected by DAMAVER were compared to each other as well as known structure of TBP, with overlays of the resulting models created with PYMOL. The fits to the experimental data of the models and the theoretical scattering of the calculated structures were generated with CRYSOL.

2.2.6.13. CLMS analysis of TAF11/TAF13/TBP complex

Cross linking mass spectrometry analysis (CLMS) was performed as described in (Rappsilber, 2011). Purified TAF11/TAF13/TBP and TAF11/TAF13 complexes were cross-linked separately by BS3 at complex/BS3 ratio of 1:25 [w/w] in crosslinking buffer (25mM HEPES, pH 8.0, 300 mM NaCl, 1mM DTT, 1 mM EDTA and Leupeptin, Pepstatin and complete protease inhibitor tablet) for 2 h on ice. The reaction was quenched by adding saturated ammonium bicarbonate solution followed by incubation on ice (45 min). Cross-linked samples were the further purified by injecting at SuperdexS200 10/300 at flow rate of 0.5mL/min. Peak fractions containing purified cross-linked samples were concentrated using amicon spin concentrators and separated on NuPAGE 4-12% bis-Tris gel gels. These cross-linking experiments were performed by Arturo Temblador, Berger group. The bands corresponding to cross-linked complexes were then excised, proteins were reduced, alkylated and trypsin digested following standard procedures and then used for MS analysis in collaboration with Juri Rappsilber at University of Edinburgh, UK.

2.2.7. Methods for characterization of 9TAF complex

2.2.7.1. 9TAF sample and negative stain grid preparation for EM

9TAF was produced as described previously (Yan Nie thesis, 2012) and purified by SEC on Superose6 10/300. 9TAF was further purified and mildly cross-linked by glutaradehyde using GraFix method (Kastner et al., 2008; Stark, 2010). The purified and cross-linked 9TAF from GraFix was subjected again to Superose6 10/300 and the peak fractions were used to prepare negative stain (2% uranyl acetate) EM grids. Sample and grid preparation was carried out by Yan Nie and Christoph Bieniossek in the Berger group. 9TAF negative stain EM RCT data was collected at Biotwin Ice CM120 Philips, EMBL-Heidelberg by Yan Nie and Xiaokang Zhang, Berger Group. 203 tilt pairs (45° tilt angle) and 168 untilted micrographs were collected at a magnification of 53,000 x (2.3 Å/pixel).

2.2.7.2. RCT (Random Conical Tilt) data collection of 9TAF and 3D reconstruction

I processed this RCT data to obtain a low resolution EM model for 9TAF. Micrographs were first preprocessed by IMOD and Bsoft and then evaluated by contrast transfer function (CTF) estimation using XMIPP software packages. 10313 particle pairs were manually picked with TiltPicker. These particle pairs were extracted from micrographs using SPIDER and XMIPP and then preprocessed using XMIPP. The untilted views of these particle pairs were used to generate 250 2D classes using IMAGIC. These 2D classes were used to generate 25 different RCT 3D models using XMIPP, out of which 4 most crowded models were selected and further refined using XMIPP. A combined pool of untilted particles from particle pair 11958 additionally picked particles from untitlted micrographs was used for this refinement. Refined models were reprojected using SPIDER for validation.

PUBLICATION 2

The Production of Multiprotein Complexes in Insect Cells Using the Baculovirus Expression System.

Authors: *Wassim Abdulrahman, Laura Radu, Frederic Garzoni, Olga Kolesnikova,* ***Kapil Gupta****, Judit Osz-Papai, Imre Berger and Arnaud Poterszman*

Methods in Molecular Biology; Volume 1261, 2015, Pages 91-114
DOI: 10.1007/978-1-4939-2230-7_5

The following manuscript explains the details of production of multiprotein complexes in insect cells using Baculovirus Expression Vector System (BEVS) such as Bac to Bac, MultiBac.

Chapter 5

The Production of Multiprotein Complexes in Insect Cells Using the Baculovirus Expression System

Wassim Abdulrahman, Laura Radu, Frederic Garzoni, Olga Kolesnikova, Kapil Gupta, Judit Osz-Papai, Imre Berger, and Arnaud Poterszman

Abstract

The production of a homogeneous protein sample in sufficient quantities is an essential prerequisite not only for structural investigations but represents also a rate-limiting step for many functional studies. In the cell, a large fraction of eukaryotic proteins exists as large multicomponent assemblies with many subunits, which act in concert to catalyze specific activities. Many of these complexes cannot be obtained from endogenous source material, so recombinant expression and reconstitution are then required to overcome this bottleneck. This chapter describes current strategies and protocols for the efficient production of multiprotein complexes in large quantities and of high quality, using the baculovirus/insect cell expression system.

Key words Recombinant protein complex production, BEVS, Baculovirus, Insect cells, Infection and coinfection methods, Multigene expression, Multiprotein complex

1 Introduction

Most eukaryotic proteins form transiently or stably interlocking assemblies that often contain many subunits. Obtaining these assemblies in a purified form in high quality and quantity is crucial for research aimed at understanding how these protein machines function in a physiological context as well as for drug discovery applications. In the absence of their interacting partners, proteins are often insoluble, improperly folded, or non-functional. Furthermore, protein complex composition and activity may vary depending on tissue type, cell state, and also on modifications of the subunits (e.g., phosphorylation, acetylation, and methylation). Production of multiprotein complexes of higher eukaryotes, including those of human origin, in a suitable form to study their structure and function can be extremely challenging. The low natural

Raymond J. Owens (ed.), *Structural Proteomics: High-Throughput Methods*, Methods in Molecular Biology, vol. 1261, DOI 10.1007/978-1-4939-2230-7_5,

abundance and heterogeneity of native complexes in cells often prevent the extraction and use of endogenous sources for the purification of protein complexes.

Recombinant approaches have become the method of choice for the production of stable macromolecular assemblies. Certain complexes may be reconstituted from recombinant proteins produced separately. Such reconstitution methods are relatively simple and particularly useful when one component of the complex is not a protein (e.g., DNA or RNA) or when the complex is short-lived or transient and cannot be purified intact. In many cases, however, this strategy is not applicable as individual subunits of a complex often cannot be expressed and manipulated in the absence of their natural partners. Multicomponent systems, including not only self-assembling multiprotein complexes but also proteins requiring specific chaperones to assist folding or post-translational modifications crucial for biological activity, will then require coexpression of a number of proteins. Many examples have shown that the simultaneous expression of different subunits of a protein complex facilitates their folding, promotes solubility, and limits degradation of regions that fold upon binding.

A number of coexpression systems are available based on *Escherichia coli* (*E. coli*) (*see* Chapter 6), mammalian cells (*see* Chapter 8), and insect cells using baculovirus vectors. Although *E. coli* is robust and inexpensive as a host, there are a number of limitations in using bacteria for synthesis of eukaryotic proteins. In particular, bacteria are unable to provide post-translational modifications and folding aids such as chaperones required for the generation of fully functional eukaryotic proteins [1, 2]. In contrast to *E. coli*, insect cells and mammalian cells have the machinery for proper folding, post-translational modification, authentic processing, and correct targeting of expressed proteins [3–5]. Several developments have been made to the baculovirus expression system, which make this the system of choice for the expression of multiprotein eukaryotic complexes. These include streamlining the assembly of multigene vectors and engineering the baculovirus genome to optimize the expression levels (for example, the MultiBac system [6–8]).

In this chapter, we describe protocols for the production of multiprotein complexes in insect cells using baculovirus expression vector systems. We describe standard procedures for working with the baculovirus system and detail the two main strategies for coexpression of multiple proteins in the same cell, which are (1) coinfection of insect cells with several viruses, each expressing a single protein and (2) infection with one single baculovirus containing all heterologous genes of choice 9.

2 Materials

2.1 Cell Culture

1. Plasticware for monolayer culture of insect cells: 25 cm^2 tissue culture flasks, six-well tissue plate with flat bottom, low evaporation lid, petri Style tissue culture dish (D60 × H15 mm) with 2 mm grid.
2. Plasticware for suspension culture of insect cells: 50 ml polypropylene tube with filter cap for oxygenation, glass or disposable Erlenmeyer flasks in different sizes.
3. Plate sealer, breathable, gas permeable, 80 × 150 mm.
4. *Sf*9, *Sf*21, and High Five cells adapted for suspension growth.
5. TNM-FH and serum-free insect cell medium.
6. Foetal Bovine Serum (FBS).
7. 1.3× SF900 medium for plaque assay.
8. Cell culture grade DMSO.
9. Cell culture grade BSA.
10. Insect cell freezing solution: 90 % insect cell medium, 10 g/L BSA, 10 % (v/v) DMSO. Sterilize solution by filtration.
11. Trypan blue: 0.4 % solution in PBS, pH 7.2.
12. Temperature controlled room or incubator set at 27 °C.
13. Platform for spinner flask operating at 27 °C and stirring up to 150 rpm.
14. Orbital shaker fitted for 250 ml to 2 l Erlenmeyer flasks, with shaking speed of up to 150 rpm (125 mm orbital). For cultures in 50 ml polypropylene tubes shake up to 250 rpm.
15. Inverted phase-contrast microscope or optionally fluorescence microscope.
16. Cell-counting chamber or optionally automated cell counter.
17. Centrifuge with adaptors for 1 l, 250 ml, 50 ml, and 15 ml tubes.

2.2 Molecular Biology

1. Commercial DNA purification kits for small scale and large scale DNA preparation.
2. 3 M Na acetate, pH 5.2.
3. T4 DNA ligase.
4. LB medium, LB agar medium.
5. IPTG, 1 M stock solution in water.
6. X-gal, 100 mg/ml stock solution in dimethylformamide.
7. Bsu36I restriction enzyme.
8. Low salt LB (for 100 ml: 1.0 g Bacto-Tryptone, 0.5 g Bacto-yeast extract, 0.5 g NaCl).

Table 1
Reagents required to generate recombinant viruses

	Tn7 transposition-based system (Method 1) (Bac-to-Bac, Multibac)	Homologous recombination (Method 2)
Method	Transfection of recombinant viral DNA	Co-transfection of linearized viral DNA with a transfer vector
Viral DNA	DH10Bac[a], DH10MultiBac, DH10EMBacY [6, 7]	BAC10:KO1629 in *E. coli* DH10B [13]
Transfer vectors (acceptors)	pFastBac[a] pFastBac Dual[a], pKL, pFL [6] pACEBac1, pACEBac2 [15]	pBacPAK8[b], pAC8 [11] pACAB3, PACAB4 [10] pAC8_DsRed[c], pAC8_MF[c]
Transfer vectors (donors)	pSPL, pUCDM, pIKD, pIDK, pIDC, … [6, 15]	
Transfer vector (polyprotein)	pPBac, pKL-pBac [12]	

[a]Invitrogen™
[b]Clontech™
[c]Available on request; (Kolesnikova et al. in prep)

9. L-Arabinose.
10. *E. coli* DH5α and TOP10 strain.
11. pBAD-His-Cre plasmid [7].
12. Tetracycline, 15 mg/ml stock solution in ethanol.
13. Kanamycin, 50 mg/ml stock solution in water.
14. Gentamicin, 10 mg/ml 7 mg/ml stock solution in water.
15. Chloramphenicol, 34 mg/ml stock solution in ethanol.
16. Zeocin, 10 mg/ml stock solution in water.

Reagents to generate recombinant baculovirus are listed in Table 1. Complementary set of oligonucleotides that correspond to the LoxP site:

Fw: 5′: ATAACTTCGTATA GCATACAT TATACGAAGTTAT 3′

Rev: 5′: ATAACTTCGTATA ATGTATGC TATACGAAGTTAT 3′

2.3 Other Buffers

1. Cre purification buffer: 20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 % glycerol, 5 mM imidazole, 5 mM DTT.
2. 10× Cre-lox reaction buffer: 0.5 M Tris–HCl (pH 7.5), 0.33 M NaCl, 0.1 M $MgCl_2$.
3. Lysis buffer: 20 mM Tris–HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40, containing protease inhibitor cocktail (PIC) and type 1 DNAase.

4. Wash buffer: 20 mM Tris–HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40.
5. Elution buffer: 20 mM Tris–HCl (pH 7.5), 250 mM NaCl, 1 mM DTT, 0.1 % NP40 with appropriate elution agent.

3 Methods

The simultaneous production of several proteins in insect cells using the BEVS requires the delivery of various genes, either by a number of individual baculoviruses (coinfection, Fig. 1a–c) or by employing a single virus that contains several genes (multigene virus, Fig. 1d–f). Coinfection enables exploratory screening of

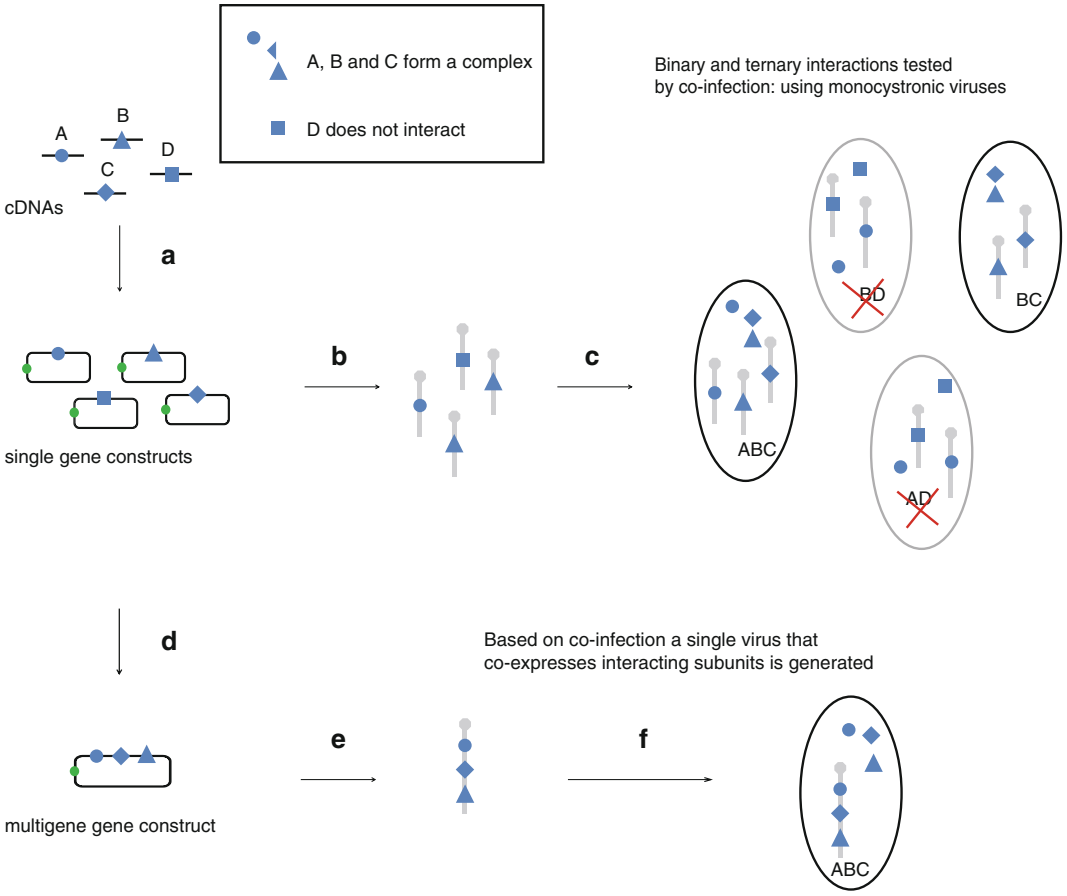


Fig. 1 Strategy for reconstitution of multiprotein complexes: (**a**) cDNAs of potential targets are cloned into transfer vectors and (**b**) the corresponding single-gene baculoviruses (expressing subunits A, B, C, and D, for example) are generated. (**c**) Insect cells are coinfected by two or three single-gene baculoviruses and association between subunits are tested by pull down, for example. (**d**) Genes encoding proteins that form stable complex (subunits A, B, and C in this example) are assembled in a single multigene transfer vector (**e**) to generate a recombinant multigene baculovirus. (**f**) Insect cells infected by multigene baculoviruses express the identified multiprotein complex A–B–C

putative interaction partners prior to large-scale expression but necessitates the maintenance of many viruses at known titers. The use of multigene baculoviruses ensures that all proteins necessary for the formation of the recombinant complex are expressed in the same infected cell, which greatly simplifies the management of the experiment.

Coexpression of multiple genes often requires extensive screening efforts to identify suitable constructs: mutations and/or deletions to optimize expression and solubility, nature and position of the affinity tag to facilitate purification. In the absence of prior knowledge about the proteins of interest, a first set of experiments with viruses expressing a single protein will provide valuable information on expression level/solubility of individual subunits. At this stage, different affinity tags can be tested [11]. For routine productions and for production of multiprotein complexes composed of a larger number of subunits, it is however preferable to use a single virus that coexpresses the different proteins. To obtain this virus, based on initial experiments, validated single gene expression units should be assembled into multigene expression cassettes using restriction/ligation or sequence and ligation independent cloning (SLIC) methods. Cre-mediated fusion of acceptor vectors with specific donor plasmids as well as the use of polyproteins for balancing subunit stoichiometry offer further options to efficiently coexpress a large number of genes [7, 8, 15].

3.1 Insect Cell Management

Working with baculoviruses requires a basic knowledge of general cell culture methods and insect cell physiology. Insect cells and viruses are handled in a laminar flow hood under aseptic conditions preferentially in absence of antibiotics, as these can mask low levels of contamination (*see* **Note 1**). All cell culture experiments are carried out at 27 °C, either in incubators or ideally in a room conditioned at this temperature. Doubling rate of the majority of insect cells (i.e., the time while the number of cells per volume increases by a factor of 2) at this temperature is around 18–20 h. The density of insect cells should range between 0.5 and 2.0×10^6 cells/ml, especially during expression experiments.

1. Remove vial of cells from liquid nitrogen and place in water bath at 37 °C. Thaw rapidly with gentle agitation until cells are almost thawed and remove cells from the water bath. Leaving cells at 37 °C after they have thawed will result in cell death.
2. Decontaminate the outside of the vial by spraying with 70 % ethanol, dry and place on ice.
3. Pre-wet a 25 cm^2 tissue culture flask by coating the adherent surface with 4 ml media.
4. Transfer the 1 ml thawed cell suspension directly into the 4 ml of media.

5. Incubate flask at 27 °C and allow cells to attach for 30–45 min.
6. After cells are attached, gently remove the medium (as soon as possible to remove the freezing solution containing DMSO).
7. Feed cells with 5 ml of fresh medium.
8. After 24 h, change to fresh medium.
9. Allow cells to grow until 90 % confluence. Detach cells by tapping the flask or by sloughing (streaming medium over the monolayer with a pipette to dislodge cells) and initiate suspension culture (Subheading 3.1, **step 2**).
10. Take an aliquot of the cell suspension (Subheading 3.1, **step 1**), count cells and determine their viability (*see* **Note 2**).
11. Add an appropriate volume of medium to a sterile Erlenmeyer, inoculate with cells to obtain a starting density of 0.5×10^6 cells/ml and incubate cells at 27 °C with agitation (80–100 rpm).
12. Monitor culture daily until cell density reaches $2–3 \times 10^6$ cells/ml and seed a fresh Erlenmeyer as in **step 2**.
13. Count cells and ensure that you have enough cells for preparing 2–4 vials (Table 2).
14. Prepare cryovials, cool them on ice.
15. Centrifuge cells at $100–150 \times g$ for 10 min at RT and remove supernatant. If High Five cells are used, make sure to keep the conditioned media when preparing freezing media.
16. Resuspend cells at the density indicated in Table 2 in the proper media.
17. Transfer 1 ml to each sterile cryovial.
18. Place at −20 °C for 1 h, then store at −80 °C for 24–48 h.
19. Transfer to dewars filled with liquid nitrogen for long-term storage.

Table 2
Media composition for freezing most commonly used cell lines

Cell line	Freezing media	Cell density (cells/ml)
*Sf*21, *Sf*9	60 % *Sf*900 medium (GIBCO) 30 % Fetal Bovine Serum (FBS) 10 % DMSO	1×10^7
High Five	42.5 % conditioned Express5 medium 42.5 % fresh Express5 medium 5 % FBS 10 % DMSO	3×10^6

3.2 Generation of Recombinant Baculoviruses

Two methods are available for the generation of recombinant baculoviruses; both make use of the baculovirus genome engineered into a bacmid for propagation in *E. coli*. The first method is based on site-specific transposition (Tn7 transposition) of an expression cassette into the baculovirus genome in *E. coli*. The second employs a transfer vector and viral DNA that are cotransfected into insect cells and utilize host enzyme-mediated homologous recombination.

3.2.1 Method 1: Tn7 Transposition and Preparation of Bacmid for Transfection

1. *Day 1*: transform 50–100 μl chemical-competent cell solution (DH10Bac, DH10MultiBac) with 10–100 ng of appropriate transfer vector, resuspend in 600 μl of LB (or SOC as preferred), and incubate cell solution at 37 °C overnight (8 h).
2. *Day 2*: streak out 150 μl on selection plates for blue/white screening containing the relevant antibiotics (Kanamycin at 50 μg/ml, Tetracyclin at 10 μg/ml, and Gentamicin at 7 μg/ml), IPTG (1 mM), and X-gal (100 μg/ml). Incubate plates at 37 °C until blue or white colored colonies can be unambiguously seen. Use dilution series (1:1, 1:10, 1:100, 1:1,000) in order to obtain optimal separation of colonies on one of the plates.
3. *Day 3*: restreak several white colonies and one blue colony as a control on the same plate in **step 2** for each construct to confirm the color of colonies.
4. *Day 4*: pick two white colonies for each construct, start 2 ml minicultures (in LB with appropriate antibiotics) overnight at 37 °C. Note that pellets can be, at this step, frozen at −20 °C for short-term storage, or kept as glycerol stock. Avoid long-term of purified bacmid (at 4 °C or −20 °C).
5. Prepare bacmid DNA using a standard plasmid purification kit, taking care not to vortex the sample during the DNA preparation to avoid shearing of the bacmid DNA. At this stage, the bacmid preparation can be checked on an agarose gel and transposition analyzed by PCR and/or sequencing (*see* **Note 3**). The recombinant bacmid is now ready for transfection into insect cells to generate the baculovirus.

3.2.2 Method 2: Preparation of Linearized Viral DNA and Transfer Vector for Cotransfection

1. *Day 1*: inoculate 50 ml of LB Broth medium starter culture supplemented with kanamycin at 50 μg/ml and chloramphenicol at 25 μg/ml with a glycerol stock of Bac10:KO1629 [13] (*see* **Note 4**). At the end of the day, use this culture to inoculate 1 l of LB Broth medium supplemented with kanamycin and chloramphenicol, and let the culture grow overnight at 37 °C.
2. *Day 2*: purify the bacmid using a large-scale DNA purification kit (Maxi prep). Follow manufacturer's protocol until DNA

elution and isopropanol precipitation. DNA pellet is then washed with 10 ml of 70 % ethanol.

3. Remove ethanol and air-dry the precipitated bacmid under the sterile hood. Resuspend DNA in 200 µl of sterile ultrapure water.
4. Transfer 10 µl of suspension to sterile 1.5 ml tube to estimate DNA concentration and to analyze. We usually obtain 20–40 µg of purified bacmid from 1 l of *E. coli* culture.
5. Linearize the bacmid with the restriction enzyme Bsu36I. For the digestion of 20 µg bacmid, mix 200 µl of bacmid (100 µg/mL), 30 µl of 10× NEB3 buffer, 3 µl of 100× BSA (optional), 57 µl of ultrapure water, and 10 µl of Bsu36I (NEB) (20 U/µl).
6. Incubate for 5 h at 37 °C, then analyze an aliquot by gel electrophoresis on a 0.8 % agarose gel to check digestion before heat inactivation (20 min at 72 °C) of Bus36I.
7. The linearized bacmid can be stored at 4 °C for 3–6 months. Alternatively, prepare aliquots of 6.5 µg (65 µl); each aliquot is sufficient for 6–12 small-scale transfections (six-well plate format) and freeze at −20 °C. Once an aliquot was thawed, keep DNA at 4 °C and do not refreeze.
8. Perform a small or medium scale DNA preparation of a suitable transfer plasmid and precipitate 10 µg of DNA with 300 mM Na-acetate pH 5.2 (final concentration) and three volumes of ethanol 100 %. Place at −80 °C for more than 1 h and centrifuge at 25,000×*g* for 15 min. Carefully remove the supernatant, add 1 ml of cold 70 % ethanol, and centrifuge it again.
9. Remove ethanol and air-dry the precipitated DNA under the sterile hood. Resuspend DNA in 20 µl of sterile ultrapure water. Take an aliquot to measure the DNA concentration and store at −20 °C.

3.2.3 Transfection of Insect Cells to Generate Recombinant Baculovirus

1. Seed 0.5–1 × 10^6 cells from a stock culture into the wells of six-well tissue culture plates. Add medium in each well to a total volume of 3 ml. In a typical experiment, include one well that contains only cells, one well that contains only medium as well as a positive fluorescent transfection control (PC), e.g., with a DsRed expression plasmid and bacmid. For each DNA construct, seed two wells.
2. For each construct, prepare transfection complexes by diluting the purified bacmid (method 1, Subheading 3.2.1) or a mixture of 0.5–1.0 µg of linearized viral DNA and 2 µg of transfer plasmid (method 2, Subheading 3.2.2) in 200 µl insect cell medium.
3. Mix 100 µl of insect cell medium with 2–10 µl of transfection reagent (e.g., Cellfectin II Life technologies) in a separate Eppendorf tube (*see* **Note 5**).

4. Add the diluted transfection reagent to the DNA solution (respect the addition order), vortex for 10 s, and incubate at room temperature (RT) for 20 min.
5. Add 1 ml of medium to the transfectant-DNA suspension and use it to replace supernatant from seeded cells. Incubate for 5 h at 27 °C, aspirate the suspension, add 3 ml of fresh medium, and return to 27 °C.
6. Incubate for a maximum of 5–6 days at 27 °C. When more than 50 % of cells in the positive control express the fluorescent protein, carefully collect the supernatant and store it at 4 °C, protected from light. This is the passage 0 (P0) virus stock. The P0 stock can be used directly to test expression of the desired proteins in small-scale experiments or has to be concomitantly amplified for production.

3.3 Virus Amplification and Storage

3.3.1 Amplification

1. Prepare 100 ml Erlenmeyer flask containing 25 ml of *Sf*9 or *Sf*21 suspension at a density of 0.5×10^6 cells/ml in exponential growth phase, infect with a small volume of P0 virus, and incubate at 27 °C with agitation (100 rpm). The volume of virus depends on the titer of your virus stock and thus on transfection/cotransfection efficiency (*see* **Note 6**).
2. Count cells 1 day postinfection and measure their size distribution. If the volume of virus added to the culture was adequate (MOI of 0.5 or below), cells should look healthy and should have doubled. At this time, infected cells should be releasing budded virus into the medium to infect other cells. If too much virus was added, you should see signs of infection: cells swell (size can increase up to 20–30 %), stop dividing, and appear uniformly rounded with enlarged nuclei. Restart with less virus.
3. Count cells 2 days postinfection and measure their size distribution. Most, if not all, of the cells should show substantial swelling, and proliferation arrest should be observed. The cell count number should be substantially below that expected if they had continued doubling every 24 h. Return the flask to a 27 °C shaker for 24 h.
4. Count cells again 24 h later and estimate their size.
5. If substantial swelling and proliferation arrest was already observed the day before, there should be no increase in cell number. Go to **step 7**.
6. If not, return the flask to a 27 °C shaker for another 24 h until proliferation arrest is determined. Dilute culture (and split in fresh flask if necessary) to maintain cell density below 2×10^6 cells/ml to prevent oxygen deprivation and entry into stationary phase. If cells do not stop doubling after 5 days, we recommend repeating the initial virus preparation.

7. The culture should be harvested when cells have been infected for about 48 h, i.e., 24 h after cells have stopped dividing. Transfer the suspension into a fresh 50 ml tube and spin for 5 min at 4,000 × *g* in a table top centrifuge. Collect the supernatant into a fresh 50 mL tube and supplement with 10 % FBS if a serum-free medium was used and store at 4 °C protected from light. This is the P1 virus stock.
8. Gently resuspend the pellet in fresh medium (respecting the cell density) and place back suspension into the shaker flask for further analysis.
9. P1 is sufficient for initial protein expression studies. If large volumes of virus are required, repeat the cotransfection or amplify P1 to obtain P2 (and eventually P3) (*see* **Note 7**).
10. Detailed monitoring of virus amplification is not always possible when a large number of viruses are needed simultaneously. Simplified protocols are described below:

Alternative protocol 1 (*5–7 days*): Infect a 250 ml suspension culture in exponential growth phase at 0.5–1 × 10^6 cells/ml by adding 0.25 ml of P0 stock and incubate under agitation at 27 °C. After 5–7 days, observe cells for signs of infection, spin the culture, and harvest the supernatant (P1 stock).

Alternative protocol 2 (*3 days+ 3 days*): Infect a 25 ml suspension culture in exponential growth phase at 0.5–1 × 10^6 cells/ml by adding 2.0 ml of P0 stock and incubate with agitation at 27 °C for 3 days. Spin the culture and harvest the supernatant (P1 stock).

Infect a 250 ml suspension culture in exponential growth phase at 0.5–1 × 10^6 cells/ml by adding 2.5 ml of P1 stock and incubate under agitation at 27 °C for 3 days. Spin the culture and harvest the supernatant (P2 stock).

3.3.2 Storing Viruses (BIIC)

Safe storage of valuable virus stocks for future use is often of paramount importance. Virus stocks can be stored at 4 °C in the dark (after addition of 10 % FBS or 0.1–1 % BSA) for 1–12 months. We found that storing virus at liquid nitrogen temperatures in frozen aliquots of baculovirus-infected insect cells (BIICs) is most advantageous in terms of virus stability and storage space requirements. Frozen aliquots of infected insect cells can be prepared and rethawed and used for protein expression without loss after extended storage times (several years). Integrity of recombinant virus can be checked by PCR (*see* **Note 8**).

1. Prepare 50 ml culture of *Sf*9 or *Sf*21 cells in exponential growth phase at 1 × 10^6 cells/ml.
2. Infect cells with a chosen volume of P1 virus.

3. Maintain cells to a concentration of 1×10^6 cells/ml until proliferation arrest.
4. Centrifuge cell culture in sterile 50 ml tube at 100–150 × *g* for 10 min, remove supernatant.
5. Resuspend cells gently in sterile freezing solution to a final density of 1×10^7 cells/ml.
6. Transfer 1 ml aliquots into sterile cryovials.
7. Place at –20 °C for 1 h.
8. Store at –80 °C for 24–48 h.
9. Store cryovials in liquid nitrogen for indefinite time.

3.4 Protein–Protein Interaction Screening by Coinfection

A set of viruses for expression of all components of the multiprotein complex are generated as described above. In the case of large subunits, this could also include subdomain constructs designed from the analysis of multiple sequence alignments, predictions of secondary structure, and disordered regions. Sequence tags for detection/purification are incorporated at either the amino terminus or carboxy terminus of the constructs, e.g., His-tag, StrepII tag. Using different tags for each component enables the expression of each to be followed, though in practice depending on the number of subunits in the complex, this may be limited by the availability of tags.

3.4.1 Construction Transfer Vectors Expressing Individual Subunits

1. Compile the cDNAs encoding the subunit sequences of interest and select a set of transfer vectors (Fig. 2a) that will be used to generate the first ensemble of recombinant viruses. Bear in mind that in a second set of experiments, you might have to generate multigene expression constructs and use an adapted transfer vector harboring a multiplication module and/or a LoxP site from the MultiBac suite to facilitate further DNA manipulations (*see* Subheading 3.5).
2. Design PCR primers for insertion of the genes into the selected transfer vectors using either ligation independent cloning technology (e.g., SLIC) or conventional restriction enzyme/ligation based method. If the transfer vectors do not encode the desired affinity/epitope tags, use nested or assembly PCR to incorporate the required nucleotide sequences.
3. For ligation independent cloning, we recommend to design primers with 20–30 bp of homology to the destination vector and to linearize the acceptor vector with two restriction enzymes (for *in silico* design web tools, *see* **Note 9**).
4. Generate recombinant baculoviruses and amplify to 100 ml stock as described in Subheading 3.3.

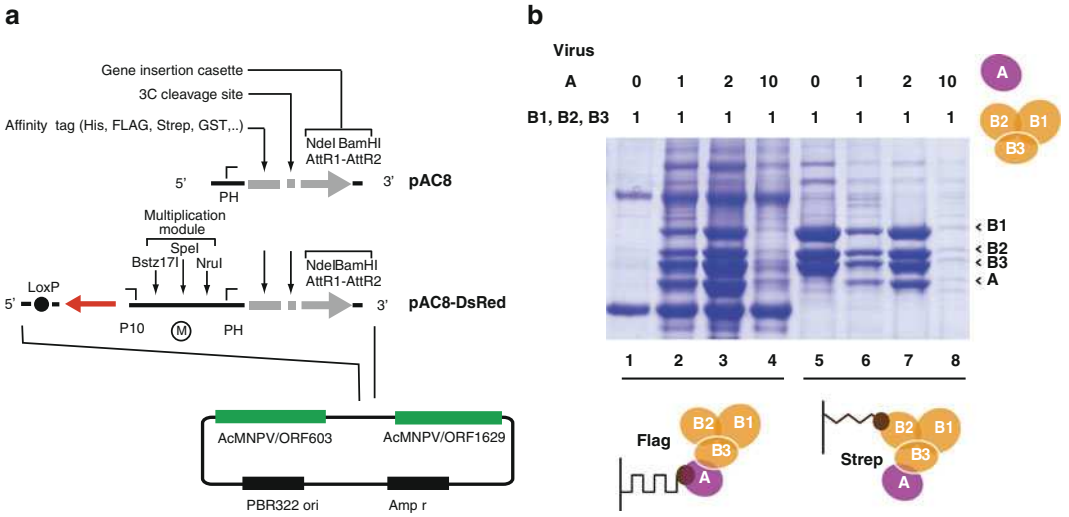


Fig. 2 Interaction screening. (**a**) pAC8 baculovirus expression shuttle vectors are tailored for interaction screening and multiprotein complex production in insect cells. This set is derived from pBacPAK8 (BD Biosciences), a shuttle vector that is used to obtain recombinant virus by homologous recombination in insect cells (homology regions are in *green*) and express the gene of interest under the dependence of the polyhedrin promoter (PH). The tag sequence is followed by a precision protease (3C) cleavage site and a gene insertion cassette (pair of unique restriction sites (NdeI and BamHI) or a Gateway cloning cassette) [9]. pAC8-DsRed derivatives also harbor a DsRed expression cassette placed under control of P10 promoter to monitor infection, a multiplication module (M) to insert an additional expression cassette as well as a loxP sequence for Cre-mediated vector fusion. (**b**) Insect cells were coinfected with baculovirus A (for expression of protein A fused with an N-terminal Flag tag) and baculovirus B1–B2–B3 (for expression of proteins B1, B2, and B3; B2 harbors strep-tag). Cells were infected with the same amount of virus A (corresponding to an MOI of 1) and increasing volumes of virus B1–B2–B3 (MOIs of 1, 2, and 10). After cell lysis, equal parts of clarified extract were mixed with anti-flag M2 beads (lanes 1–4) or streptactin-coated beads (lanes 5–8). After extensive washing, SDS loading dye was added to the beads, and samples were analyzed by SDS-PAGE. Anti-flag light and heavy chains are indicated by *black spheres*. Proteins are indicated by *arrowheads*

3.4.2 Small-Scale Expression Test

Small-scale expression tests are carried out to optimize protein production for each component of a complex individually (if expressed). Key parameters are the amount of virus and the time of infection: (1) when infecting cells for protein production, the objective is to get all cells infected synchronously. Typically, conditions that correspond to MOIs in the range of 0.5–10 are tested. (2) The best time to harvest depends on the nature of the target protein. Cells are usually analyzed 48, 72, and 96 h postinfection. Some stable proteins might accumulate to high levels 72 or 96 h postinfection while others, sensitive to degradation, will need to be collected after 24 h or most commonly 48 h. Protein expression may also depend on cell type, so expression using *Sf*9, *Sf*21, or High Five cells has to be tested.

Next, physical interactions between proteins are mapped via coinfection with multiple baculoviruses. Virus stocks from two or more viruses, each expressing a single protein (or a defined set of

proteins known to interact), are used to coinfect insect cells and thus coexpress the proteins of interests. Extracts are subjected to small-scale affinity purifications to identify protein pairs/complexes that tightly interact. There are no generic takings to determine the optimum ratio of viruses for coinfection. Optimized MOIs determined for each individual protein (Subheading 3.4.2) are a good starting point for coinfection but these conditions need to be reoptimized and several MOI combinations have to be tested. Special attention should be given to the fact that varying the MOI ratio of infecting viruses has major impact on protein expression of individual subunits and that unfavorable settings can lead to a significant decrease of the global protein production yield (Fig. 2b) see 15.

We describe a protocol for small-scale optimization of protein expression in suspension cultures in 125 ml Erlenmeyer flasks by infection with one or more baculoviruses. Optimized conditions can be scaled up to 2 and 5 l flasks for medium/large scale productions can be compared.

1. Dispense 1×10^6 cells (*Sf*9, *Sf*21, or High Five) from a culture in exponential growth phase into polypropylene tubes equipped with a filter cap for oxygen supply.
2. Pellet cells by centrifugation at $200 \times g$ for 10 min and discard the supernatant.
3. Add the desired volume of each viral stock to the cell pellet and incubate at 27 °C for 1 h with gentle agitation. Different ratios can be tested for each virus (Fig. 2b).
4. Resuspend cells in fresh medium at 1×10^6 cells/ml and incubate at 27 °C for either 48 or 72 h after infection.
5. Centrifuge the cell suspension at $200 \times g$ for 10 min in 50 ml tubes, resuspend cells in 3 ml PBS + 10 % glycerol, centrifuge again, and store pellets at −80 °C.
6. Resuspend cells in 0.8–1.5 ml lysis buffer and break cells by sonication (*see* **Note 10**). Collect 15 µl aliquots and add 5 µl of 4× SDS loading dye (total extract).
7. Clarify the lysate by centrifugation at $6{,}500 \times g$ for 60 min at 4 °C and optionally filter the supernatant using a 0.2 µm filter plate. Take a 15 µl aliquot and add 5 µl of 4× SDS loading dye (soluble extract).
8. Incubate the soluble extract with equilibrated affinity resin at 4 °C. Use 25 µl of resin for batch purification and incubate for 15–120 min with slow end-over-end mixing. To facilitate interpretations, flag at least two different proteins with two different tags and split the extract for parallel purification (Fig. 2). Use 100 µl for spin-column or filter-based chromatography. Extended incubation is not recommended as it exposes the sample to protein degradation.

9. Wash the resin with lysis buffer without PIC and elute with 50 μl of appropriate elution buffer for batch purification or with 200 μl of elution buffer for spin-column or filter-based chromatography. Take a 15 μl aliquot from each elution and add 5 μl of 4× SDS loading dye for analysis (eluted fraction).
10. As an alternative to **step 4**, add 25 μl SDS loading dye to the beads and boil the sample during 2 min prior to SDS-PAGE analysis.

At this stage, if a suitable complex is identified, production can be optimized and/or scaled up. MOIs and virus ratios as well as cell densities at infection influence the necessary duration of the culture and deserve careful analysis. Protein expression/stability may also be affected by the cell line and expression obtained using *Sf*9, *Sf*21, or High Five cells can be compared. Storage of frozen BIIC stocks will facilitate scale-up and development of reproducible production process (Subheading 3.6.2).

3.5 Construction of Multigene Baculoviruses

A number of baculovirus transfer vectors are available to enable the generation of multigene baculoviruses. These transfer vectors include pAcAB3, pAcAB4, pAcUW51, and pFastBacDual that contains two expression cassettes (Table 1). In pFastBacDUAL, PCR fragments encoding proteins of interest can be cloned into expression cassettes driven by either the p10 or polyhedrin promoters using any of the unique restriction endonuclease sites located immediately downstream of the promoters. The MultiBac technology (Fig. 3) has introduced new transfer vectors with enhanced capabilities for DNA manipulations using a multiplication module for concatenating expression cassettes, and LoxP sites for fusing Donor and Acceptor plasmids, each containing one or several expression cassettes, by means of in vitro Cre-LoxP reaction catalyzed by Cre recombinase [6–8].

3.5.1 Modification of an Acceptor Vector

Multigene expression in insect cells using the MultiBac technology relies on tandem recombineering by SLIC and/or Cre-loxP reactions between acceptor and donor plasmids [6, 7]. Any transfer vector can be converted into an acceptor plasmid by insertion of a LoxP sequence between the two DNA elements required for transposition (Tn7R and Tn7L sequence) or for homologous recombination (Orf1629 and lef2/603 sequences). Here we describe how this modification can be achieved for the transfer vector pBacPAK8 (commercially available from Clontech) by insertion of a double-stranded oligonucleotide encoding the LoxP sequence into the unique EcoRV site. Alternatively, if a unique restriction site is not available, PCR-based approaches (such as Quikchange (Stratagene)) can be used as well.

1. Analyze the plasmid map of the transfer vector and decide where the LoxP sequence has to be inserted. Identify a unique

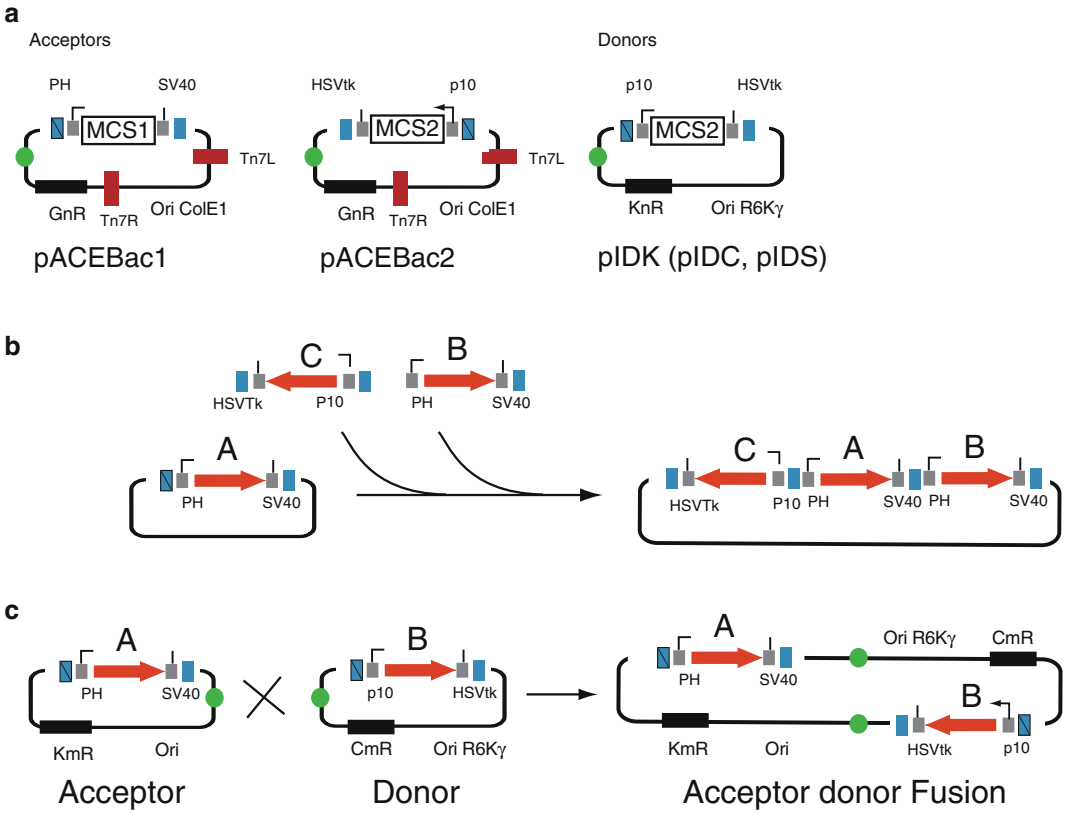


Fig. 3 MultiBac system for generation of multigene baculoviruses [6, 7, 8]. (**a**) MultiBac donor and acceptor plasmids contain multiple cloning sites (MCS) allowing insertion of your gene of interest under the control of late baculoviral promoters (PH or p10) as well as strong eukaryotic polyadenylation signal (from SV40 or HSVtk). All plasmids contain the LoxP sequence (*green filled circle*) for fusion of donors to an acceptor using Cre-mediated recombination. Acceptors have a regular origin of replication (ori ColE1); whereas, donors have a conditional origin derived from R6Kγ phage (ori R6Kγ), which allows plasmid replication exclusively in Pir-1 type *E. coli* strains. Each plasmid has a different resistance marker: gentamicin resistance (GnR) for acceptors ACEBac1 and ACEBac2, and kanamycin (KnR), chloramphenicol, or spectinomycin resistance for donors. Multiplication modules facilitating the assembly of several expression cassettes are shown as *blue boxes* flanking promoter and terminator. Acceptor plasmids contain the DNA sequences (Tn7L and Tn7R) required for transposition by the Tn7 transposase. (**b**) To assemble multigene construct using multiplication module, individual expression cassettes are excised by digestion with a pair of endonucleases and inserted via compatible restriction sites into the multiplication module of a progenitor plasmid. Following the ligation, the restriction sites used for integration are eliminated and multiplication can be repeated iteratively using multiplication module in the inserted cassette. (**c**) Acceptor and donor plasmids contain loxP sequence. Multigene constructs are assembled using Cre-mediated recombination. Resulting multigene plasmid can be propagated in a standard DH5α strain on the selective medium containing only the antibiotic resistance to which was provided by the donor vector

restriction site located within the DNA fragment that will be transferred into the baculoviral genome by transposition or homologous recombination. Do not use restriction sites located between promoters and terminators as they would interrupt expression units.

2. Digest 5 μg of plasmid with the selected restriction enzyme (EcoRV in our case), treat the plasmid with a phosphatase, isolate the linearized vector from the rest of the reaction using any purification kit available, and quantify.
3. Mix equal volume of 5′ phosphorylated complementary oligonucleotides corresponding to the LoxP site (Fwd ATAACTTCGTATAGCATACATTATACGAAGTTAT, Rev ATAACTTCGTATAATGTATGCTATACGAAGTTAT) at a final concentration of 10 μM in the presence of 5 mM $MgCl_2$. Heat to 90 °C for 2 min, ramp-cool to room temperature over a period of 45 min, and store at 4 °C.
4. Ligate 10 ng of the double-stranded LoxP fragment and 50 ng of the digested plasmid with T4 DNA ligase according to the manufacturer's protocol. Do not forget to add a negative control without LoxP fragment to evaluate the background from uncut or self-ligating plasmid.
5. Transform the ligation reaction into competent cells and plate onto LB agar containing the appropriate antibiotic.
6. Isolate colonies and purify plasmid DNA. Insertion can be checked by sequencing using a forward oligonucleotide located 150 nucleotides upstream the insertion site (*see* **Note 11**).
7. Perform an in vitro Cre-mediated fusion with a control donor vector expressing a fluorescent protein and generate the corresponding virus.

3.5.2 Cre-LoxP-Mediated Vector Fusion

A transfer vector with a LoxP site can be fused with any other plasmid that also contains a LoxP site using in vitro Cre-mediated recombination. In the MultiBac system, plasmids are divided into donors and acceptors (Fig. 3c). Acceptors contain in addition to the LoxP site, DNA elements for Tn7 transposition into bacmid-based baculovirus genomes. Donor plasmids also contain a LoxP sequence, but in contrast to the acceptor plasmids they contain a conditional R6Kγ replication origin active in *pir1* strains but not in *pir*-negative bacteria such as any commonly used cloning strains (DH5α, Top10, and others). Cre recombinase protein is commercially available but the recombinant protein can be expressed and easily purified from *E. coli* (Subheading 3.5.3).

We provide below a detailed protocol to combine an acceptor with donor(s) (Table 1).

1. Select the donor that will be used and insert the cDNA(s) of interest using your favorite technology (SLIC, RF-cloning, conventional). Don't forget to use a *pir1* strain for cloning and amplification of the plasmid.
2. Mix an equimolar amount of an acceptor and a donor plasmid (250–500 ng) with 1 μl of 10× Cre buffer and adjust the volume to 9 μl with water. Add 1 μl of Cre recombinase, mix briefly, and incubate the reaction for 30 min at 37 °C.

Avoid prolonged incubation that may lead to formation of large plasmids resulting from sequential concatenation reactions. Stop the reaction by heating at 70 °C for 15 min.

3. Transform *pir*-negative chemically competent cells (DH5α or TOP10, for example) with the Cre-treated donor–acceptor mixture. After 1 h recovery at 37 °C in absence of selection, plate cells on LB agar plates containing the antibiotic(s) to select for the presence of the donor. Selection with antibiotic corresponding to acceptor vector is not required as the propagation of the acceptor–donor fusion depends on the acceptor, which contains a common origin of replication.
4. Select 4–8 colonies. Grow 2 ml cultures in LB supplemented with appropriate antibiotics and isolate plasmid DNA using standard procedures.
5. Verify the recombination by restriction or PCR (*see* **Note 12**). Optimally, choose restriction enzyme(s) that cut in the acceptor vector and another enzyme that cuts in the vector you are inserting. This strategy will allow you to exclude multiple fusion vectors. Web tools can be used for *in silico* simulation of the Cre reaction (Cre-ACEMBLER) (*see* **Note 13**).

3.5.3 Purification of Cre Recombinase

Purified Cre recombinase can be purchased from a number of sources. Several bacterial expression plasmids are available and can be used for the production of this enzyme. Here we describe expression in *E. coli* of oligohistidine tagged Cre recombinase under the control of the arabinose promoter [7].

1. Transform DH5α or Top10 *E. coli* strain with pBADZ-His-Cre plasmid, plate onto low salt LB plates containing 25 μg/ml of Zeocin (activity of Zeocin depends on ionic strength), and incubate overnight at 37 °C.
2. Prepare a preculture in low salt LB medium containing 25 μg/ml of Zeocin that will be used to inoculate 100 mL of the same medium. The starting OD_{600nm} should be in the range 0.05–0.1. Place it in the incubator for growth.
3. When the OD_{600nm} reaches 0.5–0.6, add 1 ml of 20 % L-ARABINOSE (0.2 % final concentration) to induce the expression of the recombinant protein. Don't forget to take a 100 μl aliquot as control before induction (uninduced sample).
4. Let the culture grow for 4 h more, take a 100 μl aliquot as control (induced sample), and harvest cells by centrifugation at 4,000 × *g* for 15 min. Wash the cell pellet with PBS buffer supplemented with 10 % glycerol, centrifuge again, and store the pellet at −20 °C.
5. Add 20 μl 4× SDS loading dye to uninduced and induced sample and boil it for 5 min.

6. Check the expression on 15 % SDS gel, load 10 µl of the sample without induction and 5 µl of the sample after induction. The band corresponding to the recombinant Cre protein should be visible at 38 kDa.
7. Resuspend the stored cell pellet in 15 ml of lysis buffer (20 mM Tris–HCl (pH 8.0), 300 mM NaCl, 10 % glycerol, 5 mM imidazole, 5 mM DTT) and sonicate for 1.5 min at 40 % amplitude.
8. Centrifuge at 20,000 × *g* for 20 min at 4 °C. Keep the supernatant.
9. Wash 1 ml IMAC resin with ultrapure water then equilibrate with the lysis buffer. Always handle samples on ice or work in a cold room.
10. Add the equilibrated resin to the supernatant and incubate for 2 h at 4 °C.
11. Centrifuge at 200 × *g* for 10 min to pellet the resin. Discard the supernatant.
12. Wash the resin with 10 ml of lysis buffer containing 10 mM imidazole.
13. Centrifuge again at 200 × *g* for 10 min to pellet the resin and discard the supernatant. Repeat the washing step once more.
14. Poor the resin in an empty gravity flow column and elute with lysis buffer containing 150 mM imidazole.
15. Analyze eluted protein by 15 % of SDS gel, pool the protein containing fractions and dialyze overnight in a buffer containing 20 mM Tris–HCl (pH 8.0), 300 mM NaCl, and 20 % glycerol.
16. Measure the protein concentration and snap freeze 100 µl aliquots at 0.3 mg/ml concentration. Store at –80 °C.
17. Perform a set of in vitro Cre-mediated fusion reactions with a control vectors using increasing amounts of recombinase for quality control.
18. From 100 ml culture, you should obtain enough recombinase for 100 reactions.

3.6 Scale-Up of Protein Expression

Routine preparation of protein complexes in quantity and quality sufficient for biotechnology applications may necessitate significant efforts, in particular when large volumes of culture are required. Simple protocols to streamline large/medium-scale production are described below.

3.6.1 Use of Fluorescent Reporter Protein

Fluorescent proteins such as the yellow fluorescent protein (YFP) or the *Discosoma* red fluorescent protein (DsRed) can be efficiently used to monitor virus performance and determine the optimal time for harvesting the culture and proceeding to protein purification.

The genes coding for fluorescent proteins can be either fused as a tag to the recombinant protein of choice or integrated directly into the baculoviral genome. In the latter case, if the gene encoding the fluorescent protein is placed under the control of the same promoter that drive the expression of heterologous proteins of interest, detection of fluorescent signal might sign the concomitant expression of proteins of interest. Thus, a plateau of fluorescence is observed when production of heterologous protein is maximal. Toward this goal, the YFP gene was inserted under the control of the polyhedrin promoter into the backbone of the MultiBac virus, giving rise to the virus EMBacY [15].

To determine the optimal time for harvesting cells when protein production is maximal, fluorescence measurements of a reporter gene can be conveniently used as follows:

1. Count cells in an expression culture and withdraw 2×10^6 cells. Centrifuge for 1 min at 15,000×*g*, remove supernatant, and resuspend the pellet in 1 ml of PBS (or any buffer of choice, if for example the optimal lysis buffer has been already determined).
2. Sonicate with 3 mm probe at 20 % of intensity until the pellet is completely dissolved (do not overheat).
3. Take out 50 µl, transfer into a fresh Eppendorf tube, and add SDS-PAGE loading dye (total extract).
4. Spin down the rest of the sample (950 µl) for 3 min at maximum speed in table top centrifuge.
5. Take out 50 µl, transfer into fresh Eppendorf tube, and add SDS-PAGE loading dye (soluble extract).
6. Use remaining 950 µl to measure fluorescence (excitation: 488 nm, emission max: ~520 nm if a baculovirus expressing YPF such as EMBacY is used).
7. We strongly advice to measure the cellular fluorescence against a fluorescent standard in order to calibrate the measurements and to compare with expressions from other batches of virus.
8. Freeze the SNP/SN aliquots at –20 °C until use.

3.6.2 Using BIIC Stocks for Protein Production

(1 ml of BIIC for 1 l of expression culture.)

1. Quickly thaw one BIIC cryovial in a 37 °C water bath (or thaw in your hands, then use paper towels to protect your skin) with gentle agitation until cells are almost thawed.
2. Dilute the vial quickly into 100 ml insect cell medium.
3. Add this 100 ml suspension to 900 ml of uninfected *Sf*21 cells at a density of around 0.9×10^6 cells/ml.
4. Maintain cells at a density of 1×10^6 cells/ml until proliferation arrest.

5. Harvest cells at proper time after proliferation arrest (if EMBacY-based virus is used: when YFP reaches a plateau) and purify your protein.

4 Notes

1. Addition of penicillin (5–25 U/ml) and streptomycin (7 μg/ml) or gentamicin (5 μg/ml) can be useful when it is necessary to face a contamination.
2. Cell viability can be evaluated with Trypan blue. Mix one volume of cells with one volume of a 0.1 % stock solution of Trypan blue (in PBS or other isotonic salt solution). Nonviable cells will take up Trypan blue. Healthy cultures should contain more than 97 % of unstained viable cells. We use an automated counter that also provides cell size distribution, which is an indicator of cell infection.
3. In our experience, it is absolutely necessary to ascertain at this step that debris from the cell lysis are completely removed. Leftover debris, containing also genomic nucleic acid and RNA, can inhibit the transfection of insect cells by sequestering transfection reagent. PCR analysis can be performed with m13 Forward (CCCAGTCACGACGTTGTAAAACG) and m13 Reverse (CAGGAAACAGCTATGAC) primers. The size of amplified DNA fragment can be compared with the size of DNA fragment obtained from non-recombinant bacmid propagated by blue colonies. Column purified PCR reaction can be sequenced to check the integrity of inserted expression cassette.
4. We mainly use the bacmid BAC10:KO1629 [13] as a source of viral DNA. It consists of the wild-type AcMNPV genome in which part of ORF1629 that is essential for virus replication has been replaced by a low copy bacterial replicon and resistance makers (Kanamycin and Chloramphenicol). The bacterial replicon and resistance markers are surrounded by two Bsu36I restriction sites used for DNA linearization which increases recombination efficiency. ORF1629 is rescued after recombination with the transfer vector. Note that ready-to-use and genetically optimized linearized baculovirus DNA can be purchased from a number of sources (BaculoGold™ (BD Pharmingen), BacMagic™ and BacVector™ (Novagen), BacPAK6™ (Clontech), flashBAC™ (OET), Sapphire™ (Allele Biotechnology)).
5. Note that for initial transfection many other transfection reagents can be used as well. In our hands transfection with Fectofly (Polyplus), FuGENE, or X-Treme GENE HP (Roche) works with the same efficiency and offers the advantage that

the transfection suspension does not need to be removed after the 5 h incubation. Respect the incubation time recommended as extended incubation may lead to the formation of large and difficult to transfect DNA/transfection complexes. Carefully follow the manufacturer's instructions for procedures and the choice of compatible medium.

6. For amplification the multiplicity of infection (MOI) should be 0.1–0.4. Ideally the titer of the P0 stock should be determined experimentally. However, assuming a cell concentration of 0.5×10^6 cells/ml and a titer of the P0 stock of 2×10^7 pfu/ml, then an MOI of 0.1–0.4 would correspond to 0.25–1 % (virus volume/culture volume percentage). For example, add between 70 and 240 μl of virus stock to a 25 ml culture in 250 ml flask.
7. Defective interfering particles (DIPs) are viral particles that miss part or all of their genome and thus cannot sustain an infection by themselves. Instead, they depend on coinfection with a suitable helper virus that provides the gene functions absent from the DIPs. Accumulation of DIPs that arise during passaging is favored by amplification at high MOIs when cells can be coinfected with an intact virus and a DIP, allowing their replication. At low MOIs, formation of DIPs is limited as each cell is infected by a single viral particle. If the virus is defective, it will not replicate and DIPs will not accumulate [14, 16].
8. For PCR analysis of recombinant viruses, treat 200 μl of viral suspension with 20 μl of proteinase K at 20 mg/ml in 10 mM Tris–HCl (pH 8.0), 1 mM EDTA, 0.5 % SDS for 10 min at 72 °C and extract DNA with purification kit suitable for purification of large DNAs. This should provide enough DNA for running 100 PCR reactions.
9. For *in silico* design of oligonucleotides suitable for SLIC, Gibson or InFusion cloning uses SODA (http://slic.cgm.cnrs-gif.fr/) or NebBuilder (http://nebuilder.neb.com/).
10. One should consider factors that might affect the stability of the putative complex (pH, ionic strength, nonprotein ligand). Preparation of the cell lysate is a critical step that often requires optimization to identify a suitable lysis buffer. Optimal conditions should maximize the solubility and stability of the complex while minimizing oxidation, unwanted proteolysis, and aggregation. If the complexes can be tested in vitro, screening should include the use of a functional assay to control/optimize the activity of recombinant proteins.
11. Blunt-end ligation protocol allows insertion of the LoxP sequence in two possible orientations that will lead to distinct Cre-mediated fusion plasmids. A priori, there is no reason to privilege one orientation rather than the other, nevertheless

one cannot exclude that one orientation could be more favorable for virus stability.

12. For expression of two cDNAs, we would place one gene under the control of the PH promoter and another under that of promoter p10. For three, we would use two PH promoters and one p10 or vice versa. For expression of four cDNAs, use two PH and two p10 promoters.
13. If Cre/loxp concatenation between three plasmids (e.g., A, B, and C) does not work or interpretation is difficult (large plasmids and complex restriction patterns), proceed sequentially, i.e., isolate AB or BC fusion first and recombine with the third plasmid in a second step. Addition of Cre recombinase to a mixture of two plasmids leads to equilibrium after approximately 30 min at 37 °C. At equilibrium, only 20 % of initial plasmids are recombined (AB vector). For in silico modeling of a Cre-lox, recombination reaction uses a software Cre-ACEMBLER-1.0.1 (http://www.embl. fr/multibac/multiexpression_technologies/cre-acembler/.)

Acknowledgments

We thank Alice Aubert, Petra Drnkova, Maxime Chaillet, Isabelle Kolb, Natalie Troffer-Charlier, and Jean-Marie Garnier for sharing their experience on baculovirus expression and molecular biology. This work was funded by the CNRS, the INSERM, the Université de Strasbourg (UdS), the Alsace Region, and the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01 Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions. It benefited from grants ANR-12-BSV8-0015-01 from the Agence Nationale de la Recherche, INCA-2008-041 from the Institut National du Cancer, the Association pour la Recherche sur le Cancer, the Fondation pour la Recherche Médicale (FRM) (ING20101221017), and La Ligue contre le Cancer (fellowship to LR). IB acknowledges support from the European Commission (EC) Framework Programme (FP) 7 project ComplexINC (279039).

References

1. Baneyx F (1999) Recombinant protein expression in Escherichia coli. Curr Opin Biotechnol 10:411–421
2. Brondyk WH (2009) Selecting an appropriate method for expressing a recombinant protein. Methods Enzymol 463:131–147
3. Assenberg R, Wan P, Geisse S, Mayr L (2013) Advances in recombinant protein expression for use in pharmaceutical research. Curr Opin Struct Biol 23:393–402
4. Khan KH (2013) Gene expression in mammalian cells and its applications. Adv Pharm Bull 3:257–263

5. Picanco-Castro V, Biaggio RT, Cova DT, Swiech K (2013) Production of recombinant therapeutic proteins in human cells: current achievements and future perspectives. Protein Pept Lett 20:1373–1381
6. Fitzgerald DJ, Berger P, Schaffitzel C et al (2006) Protein complex expression by using multigene baculoviral vectors. Nat Methods 3:1021–1032
7. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22:1583–1587
8. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37:49–57
9. Sokolenko S, George S, Wagner A et al (2012) Co-expression vs. co-infection using baculovirus expression vectors in insect cell culture: benefits and drawbacks. Biotechnol Adv 30:766–781
10. Belyaev AS, Roy P (1993) Development of baculovirus triple and quadruple expression vectors: co-expression of three or four bluetongue virus proteins and the synthesis of bluetongue virus-like particles in insect cells. Nucleic Acids Res 21:1219–1223
11. Abdulrahman W, Uhring M, Kolb-Cheynelet I et al (2009) A set of baculovirus transfer vectors for screening of affinity tags and parallel expression strategies. Anal Biochem 385:383–385
12. Nie Y, Bellon-Echeverria I, Trowitzsch S et al (2014) Multiprotein complex production in insect cells by using polyproteins. Methods Mol Biol 1091:131–141
13. Zhao Y, Chapman DA, Jones IM (2003) Improving baculovirus recombination. Nucleic Acids Res 31:E6–6
14. Kool M, van Lier FLJ, Vlak JM, Tramper J (1991) Detection and analysis of Autographa californica nuclear polyhedrosis virus mutants with defective interfering properties. Virology 183:739–746
15. Vijayachandran LS, Viola C, Garzoni F et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175:198–208
16. Pijlman GP, van den Born E, Martens DE, Vlak JM (2001) Autographa californica baculoviruses with large genomic deletions are rapidly generated in infected insect cells. Virology 283:132–138

PUBLICATION 3

The MultiBac protein complex production platform at the EMBL.

Authors: *Imre Berger, Frederic Garzoni, Maxime Chaillet, Matthias Haffke,* ***Kapil Gupta*** *and Alice Aubert*

Journal of Visualized Experiments, Issue 77, November 2013, Pages 91-114
DOI: 10.3791/50159

The following video article describes about the multiprotein complex production platform at the EMBL. At this platform baculovirus based MultiBac system is used for protein expression in insect cells. Detailed protocol starting from creation of multigene expression constructs until protein production is explained to full details.

Video Article

The MultiBac Protein Complex Production Platform at the EMBL

Imre Berger[1], Frederic Garzoni[1], Maxime Chaillet[1], Matthias Haffke[1], Kapil Gupta[1], Alice Aubert[1]

[1]EMBL Grenoble Outstation and Unit of Virus Host Cell Interactions (UVHCI) UMR5322

Correspondence to: Imre Berger at iberger@embl.fr

URL: http://www.jove.com/video/50159
DOI: doi:10.3791/50159

Keywords: Molecular Biology, Issue 77, Genetics, Bioengineering, Virology, Biochemistry, Microbiology, Basic Protocols, Genomics, Proteomics, Automation, Laboratory, Biotechnology, Multiprotein Complexes, Biological Science Disciplines, Robotics, Protein complexes, multigene delivery, recombinant expression, baculovirus system, MultiBac platform, standard operating procedures (SOP), cell, culture, DNA, RNA, protein, production, sequencing

Date Published: 7/11/2013

Citation: Berger, I., Garzoni, F., Chaillet, M., Haffke, M., Gupta, K., Aubert, A. The MultiBac Protein Complex Production Platform at the EMBL. *J. Vis. Exp.* (77), e50159, doi:10.3791/50159 (2013).

Abstract

Proteomics research revealed the impressive complexity of eukaryotic proteomes in unprecedented detail. It is now a commonly accepted notion that proteins in cells mostly exist not as isolated entities but exert their biological activity in association with many other proteins, in humans ten or more, forming assembly lines in the cell for most if not all vital functions.[1,2] Knowledge of the function and architecture of these multiprotein assemblies requires their provision in superior quality and sufficient quantity for detailed analysis. The paucity of many protein complexes in cells, in particular in eukaryotes, prohibits their extraction from native sources, and necessitates recombinant production. The baculovirus expression vector system (BEVS) has proven to be particularly useful for producing eukaryotic proteins, the activity of which often relies on post-translational processing that other commonly used expression systems often cannot support.[3] BEVS use a recombinant baculovirus into which the gene of interest was inserted to infect insect cell cultures which in turn produce the protein of choice. MultiBac is a BEVS that has been particularly tailored for the production of eukaryotic protein complexes that contain many subunits.[4] A vital prerequisite for efficient production of proteins and their complexes are robust protocols for all steps involved in an expression experiment that ideally can be implemented as standard operating procedures (SOPs) and followed also by non-specialist users with comparative ease. The MultiBac platform at the European Molecular Biology Laboratory (EMBL) uses SOPs for all steps involved in a multiprotein complex expression experiment, starting from insertion of the genes into an engineered baculoviral genome optimized for heterologous protein production properties to small-scale analysis of the protein specimens produced.[5-8] The platform is installed in an open-access mode at EMBL Grenoble and has supported many scientists from academia and industry to accelerate protein complex research projects.

Video Link

The video component of this article can be found at http://www.jove.com/video/50159/

Introduction

Biological activity is controlled by assemblies of proteins and other biomolecules that act in concert to catalyze cellular functions. Notable examples include the machinery that transcribes the hereditary information contained in DNA into messenger RNA. In humans, more than 100 proteins come together in a defined and regulated process to transcribe genes, forming large multiprotein complexes with 10 and more subunits including RNA polymerase II and the general transcription factors such as TFIID, TFIIH and others.[9] Other examples are the ribosome, consisting of many proteins and RNA molecules, that catalyzes protein synthesis, or the nuclear pore complex which is responsible for shuttling biomolecules through the nuclear envelope in eukaryotes. A detailed architectural and biochemical dissection of essentially all multicomponent machines in the cell is vital to understand their function. The structure elucidation of prokaryotic and eukaryotic ribosomes, for instance, constituted hallmark events yielding unprecedented insight into how these macromolecular machines carry out their designated functions in the cell.[10,11]

Ribosomes can be obtained in sufficient quality and quantity for detailed study by purifying the endogenous material from cultured cells, due to the fact that up to 30% of the cellular mass consists of ribosomes. RNA polymerase II is already less abundant by orders of magnitude, and many thousand liters of yeast culture had to be processed to obtain a detailed atomic view of this essential complex central to transcription.[12] The overwhelming majority of the other essential complexes are however present in much lower amounts in native cells, and thus cannot be purified adequately from native source material. To render such complexes accessible to detailed structural and functional analysis requires heterologous production by using recombinant techniques.

Recombinant protein production had a major impact on life science research. Many proteins were produced recombinantly, and their structure and function dissected at high resolution. Structural genomics programs have taken advantage of the elucidation of the genomes of many organisms to address the gene product repertoire of entire organisms in high-throughput (HT) mode. Thousands of protein structures have thus been determined. To date, the most prolifically used system for recombinant protein production has been *E. coli*, and many expression systems

have been developed and refined over the years for heterologous production in this host. The plasmids harboring a plethora of functionalities to enable protein production in *E. coli* fill entire catalogues of commercial providers.

However, *E. coli* has certain limitations which make it unsuitable to produce many eukaryotic proteins and in particular protein complexes with many subunits. Therefore, protein production in eukaryotic hosts has become increasingly the method of choice in recent years. A particularly well-suited system to produce eukaryotic proteins is the baculovirus expression vector system (BEVS) that relies on a recombinant baculovirus carrying the heterologous genes to infect insect cell cultures cultivated in the laboratory. The MultiBac system is a more recently developed BEVS which is particularly tailored for the production of eukaryotic protein complexes with many subunits (**Figure 1**). MultiBac was first introduced in 2004.[13] Since its introduction, MultiBac has been continuously refined and stream-lined to simplify handling, improve target protein quality and generally making the system accessible to non-specialist users by designing efficient standard operating procedures (SOPs).[4] MultiBac has been implemented in many laboratories world-wide, in academia and industry. At the EMBL in Grenoble, transnational access programs were put in place by the European Commission to provide expert training at the MultiBac platform for scientists who wished to use this production system for advancing their research. The structure and function of many protein complexes that were hitherto not accessible was elucidated by using samples produced with MultiBac.[4] In the following, the essential steps of MultiBac production are summarized in protocols as they are in operation at the MultiBac facility at EMBL Grenoble.

Protocol

1. Tandem Recombineering (TR) for Creating Multigene Expression Constructs

1. **Planning the co-expression strategy**. Design approach for inserting your genes of interest into Donors and Acceptors. Potential physiological submodules of your complex should be grouped together on specific Acceptors and Donors. Use Multiplication Module consisting of Homing endonuclease (HE) - BstXI pairs to combine expression cassettes on individual Donor and Acceptor plasmids.[7,8] Create all relevant constructs *in silico* and validate strategy thoroughly before proceeding to experimental work. For example, genes of interest should be checked not to contain HE or other restriction sites and the presence of correct open reading frames (ORFs) should be validated. Consider ordering synthetic genes optimized for insect cell codon usage and mRNA secondary structure to improve protein production levels as well as removal of any existing HE sites from the genes of interest. Consider placing purification tags based on data from the literature about flexible or exposed N- or C-termini of your proteins of choice. Consider applying polyprotein strategies that aim at producing several protein subunits in your complex if the relative amounts of individual proteins need to be controlled due to stoichiometry issues in the complex.[4] Prepare detailed "How-To" document (electronic lab book is recommended) containing all projected experimental steps of the project leading up to the complete multigene construct(s). Create electronic files of the Cre-LoxP fused plasmids for example by using the Cre-ACEMBLER software which can be downloaded from the Berger group home page (www.embl.fr/multibac/multiexpression_technologies/cre-acembler).
2. Insert your genes of interest into selected Donors and Acceptors by using restriction enzymes and ligase, or, alternatively, by using ligation independent methods following published protocols.[5,6,14] If you have access to a liquid handling work-station and if you plan a large number of constructs to be generated (for example for combinatorial approaches) consider using robotics scripts developed and implemented by the Berger group (**Figure 2**).[14,15] If a liquid handling work-station is not available, manual operation using microtitre plates allows gene insertion in a HT like fashion.
3. Constraints imposed by the need to control the stoichiometry of the expressed subunits may materialize. In the case of stoichiometrically imbalanced expression levels of individual subunits of a protein complex, consider applying polyprotein strategies to conjoin several subunits of your complex and a specific protease (for example tobacco etch virus NIa protease) in single large ORFs spaced by specific proteolytic sites.[4,8] Consider co-expressing one or several polyproteins with single expression cassettes if you have a very large complex with many subunits and widely ranging molecular weights of individual subunits. Consider co-expressing several genes encoding for the same protein in a polyprotein or as several identical expression cassettes if that protein is characterized by low production yield.[4,13]
4. Validate all Donor and Acceptor constructs cloned by restriction mapping (optionally in high-throughput) and sequencing. Proceed to fuse Donor-Acceptor combinations by Cre-LoxP recombination to generate the multigene expression constructs of choice. Validate purified Acceptor-Donor fusion plasmids by restriction mapping, use electronic sequences created by Cre-ACEMBLER or similar programs as a reference.
5. Store purified and validated Donors, Acceptors and Donor-Acceptor fusions at -20 °C or -80 °C. Archive plasmids and their sequences (Microsoft Excel, Filemaker, others) carefully for later usage.

2. Composite Multigene Baculovirus Generation, Amplification and Storage

1. Integrate multigene transfer vectors the MultiBac baculoviral genome by transforming into DH10 cells harboring the viral genome and the functionalities required for Tn7 transposition. Note that the Multibac baculoviral genome can be preloaded with particular genes of interest (YFP marker, chaperones, etc.) in its own LoxP site (engineered into the genome distal to the Tn7 attachment site) by an *in vivo* Cre reaction preceding Tn7 integration.[13] After Tn7 transposition, cells with composite baculovirus containing the genes of interest are selected by blue/white screening (successful Tn7 transposition results in loss of α-complementation of the β-galactosidase; therefore, colonies with correct Tn7 transposition remain white on selective agar plates containing X-gal) and the genome is prepared by alkaline lysis and ethanol/isopropanol precipitation.[5,6]
2. **Transfection and initial virus production**. Place 6-well tissue culture plate into sterile hood. From log-phase Sf21 insect cell culture, seed out aliquots of cells in the wells and transfect by adding the purified baculoviral genome and a transfection reagent mixed in culture media as described.[6] Harvest initial virus after 48-60 hr by removing the media (high quality, low titer virus V_0, typically 3 ml per well). Supplement fresh media, and test for protein production (and, if a YFP marker is present, for fluorescence) after an additional 2-3 days.[6,7]
3. **Amplification of virus, low MOI regimen**. Use V_0 virus to infect 25-50 ml of cells in log phase (cell density <1x10^6 cells per ml) in small (100-250 ml) Erlenmeyer shaker flasks agitated on orbital platform shakers (**Figure 3**). Count cells and split every 24 hr until cells stop doubling (proliferation arrest). Follow a low MOI (multiplicity of infection *i.e.* number of virus particles per cell) regimen: Cells must double (at least) once (MOI < 1), otherwise repeat experiment with a smaller volume of V_0 added. Normally, 3 ml of V_0 are used to infect 25 ml

of Sf21 insect cells at a density of $0.5x10^6$ cells / ml. This is essential to prevent detrimental over-amplification and auto-deletion of virus that can result in loss of the heterologous genes of interest. Harvest V_1 virus (25-50 ml) after 48-60 hr by pelleting cells and removing the media containing the virus. Supplement with fresh media and test for YFP and protein production by removing $1x10^6$ cells every 12 or 24 hr, pelleting and validating protein production and marker protein (YFP) signal.[6], Amplify virus (to V_2) further if larger expression volumes are aimed at by infecting up to 400 ml cells in 2 L shaker flasks with V_1 virus respecting the above low MOI regimen (cells must double at least once after infection with V_1). Stringently test protein production and marker protein signal during amplification to avoid accumulation of defective viruses no longer containing your genes of interest.[5-7] Use cell pellets accumulating at each amplifications step already for establishing purification protocols for the expressed protein complex of interest.

4. **BIIC storage of production virus**. We strongly recommend to store V_2 virus as the production virus by using the BIIC (**B**aculovirus-**i**nfected **i**nsect **c**ells) method, to prevent modifications (*e.g.* loss of the gene of interest) of the recombinant virus and to preserve high expression levels.[16] Pellet infected cells 24 hr after proliferation arrest is observed - at this stage cells contain complete viral particles just before they would be released (by budding) into the media. Remove media and freeze aliquots of the cell pellet in liquid nitrogen and store indefinitely.[7,16]

3. Protein Production and Downstream Processing

1. **Infecting large(r) cultures and monitoring YFP**. Use V_1, V_2 or frozen BIIC aliquots to infect larger cell cultures for production runs (typically 400 ml in 2 L flasks). Adhere to low MOI regimen (adjust virus volume used for infection such that infected culture doubles at least once). Enlarge infected culture volumes if needed by multiplying number of flasks. If YFP marker protein is present, withdraw at defined intervals $1x10^6$ cells, pellet and lyse cells and monitor evolution of the YFP signal until a plateau is reached indicating maximum recombinant protein production. YFP levels can be measured in a standard 96 well plate reader capable of recording fluorescence signals (*e.g.* Tecan SPECTRAFluor). Harvest cells at this stage. Store cell pellets at -20 °C (short term) or -80 °C (long term).
2. **Cell lysis and fractionation**. Lyse cells by your favorite method of choice, tailored to the requirements of your protein (freeze-thaw, sonication, French press, others).[5-7] Fractionate cytosol and nuclei and test for the presence of your proteins of interest. Develop purification protocols based on the results to simplify protein purification. Consider applying soaking procedures to extract your protein from the nuclear fraction under high KCl conditions if your proteins reside in the nucleus.[7,18]
3. **Protein purification (micro-scale, large scale)**. Note that often small volumes (10 or 25 ml) of cell culture are sufficient for obtaining cell pellets for purifying substantial amounts of your proteins of interest due to the typically high or very high production levels of heterologous proteins in the baculovirus/insect cell systems (often 10-100 mg of protein per L culture and more). In conjunction with micro-purification (multiwell plates, microtip methods, GE Healthcare ÄKTAmicro system, others) it is possible to obtain biochemical and activity data and often also sufficient amount of the desired proteins and complexes for nanoliter-scale high-throughput crystallization (HTX). Consider using metal affinity purification (Clonetech Takara TALON, Qiagen NiNTA metal chelator resins) and an oligo-histidine (6-10 residues) tag on exposed subunits of your protein complex to facilitate purification, in conjunction with ion exchange and size exclusion chromatography in small volumes using for example the ÄKTAmicro or a similar small volume purification machine (**Figure 4**). Other affinity purification steps and ion exchange (IEX) steps in addition to the affinity purification with oligo-histidine tags or other tags (choose from GST, MBP, CBP, others) can and should be considered, according to the biochemical properties of the proteins of interest and individual preferences. Consider tagging more than one subunit with affinity tags to enhance purification efficiency. Concentrate your purified protein complexes and establish storage protocols such as freezing with or without glycerol. Develop quality control criteria that can be applied standardly (activity assays, biochemical and biophysical tests) to assess batch-to-batch variation of your purified proteins.

Representative Results

Strong co-expression of heterologous proteins achieved by the MultiBac system is shown in **Figure 1d** (probes taken 48 hr after infecting a suspension cell culture). The overexpressed protein bands are clearly discernible in the whole cell extract (SNP) and the cleared lysate (SN). The quality and quantity of the protein material produced is often sufficient to enable structure determination of protein complexes, such as the mitotic checkpoint complex MCC shown in **Figure 1e**.[17]

Figure 2 displays the work-flow of a gene assembly experiment by robot-assisted tandem recombineering (TR). Robust DNA assembly protocols were scripted into robotics routines for parallelized assembly of multigene expression constructs. Individual robotic steps are shown in snap-shots (**Figure 2c, I - IV**). The DNA components to be assembled are generated by PCR and quality controlled by e-gels (**Figure 2d**, left); the assembled multigene constructs are likewise validated by PCR with specifically designed sets of primers (**Figure 2d**, right).[14,15]

Recombinant baculovirus generation and amplification follows standard operating procedures (schematically displayed in **Figure 3a**). Snapshots of cells cultured in a monolayer are shown (**Figure 3b**, I. & II.) following infection with a MultiBac virus.

Downstream processing of the recombinant protein complexes can be miniaturized by using multi-well-plate or microtip-based chromatography for affinity purification followed by size exclusion chromatography (SEC) of the complexes by integrating into the work-flow small-scale systems such as the ÄKTAmicro (**Figure 4a**). A representative SEC profile of a ~700 kDa human transcription factor complex is shown. Sample purified by using the ÄKTAmicro in small-scale is typically sufficient for characterization by biochemical and biophysical means including electron microscopy (**Figure 4b**).

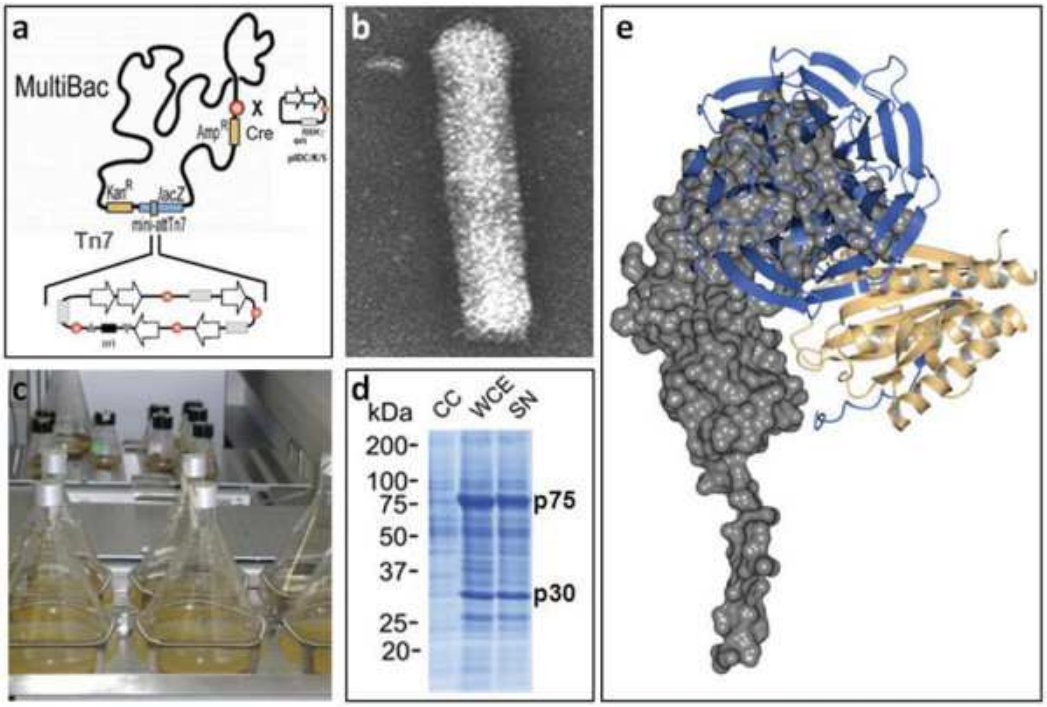


Figure 1. MultiBac platform technology for multiprotein complex production. (a) Genes of interest are integrated into the Multibac baculoviral genome by using Tn7 transposition in conjunction with blue/white screening. A LoxP site on the virus backbone allows addition of further functionalities such as a fluorescent marker protein to monitor virus performance and heterologous protein production. **(b)** The baculovirus adopts an elongated stick shape and is characterized by a flexible envelope that can augment to accommodate the large (>130 kb) circular double-stranded DNA genome. Large heterologous gene insertions in the genome are tolerated by elongating the envelope. **(c)** Standard Erlenmeyer flasks on orbital shaker platforms can be used for growing large scale insect cell cultures for heterologues protein production. **(d)** The MultiBac system efficiently produces recombinant proteins which are often clearly visible already in the whole-cell extract (SNP). **(e)** The structure of the mitotic checkpoint complex was elucidated by X-ray diffraction from crystals grown from sample produced with the MultiBac system.[17] Click here to view larger figure.

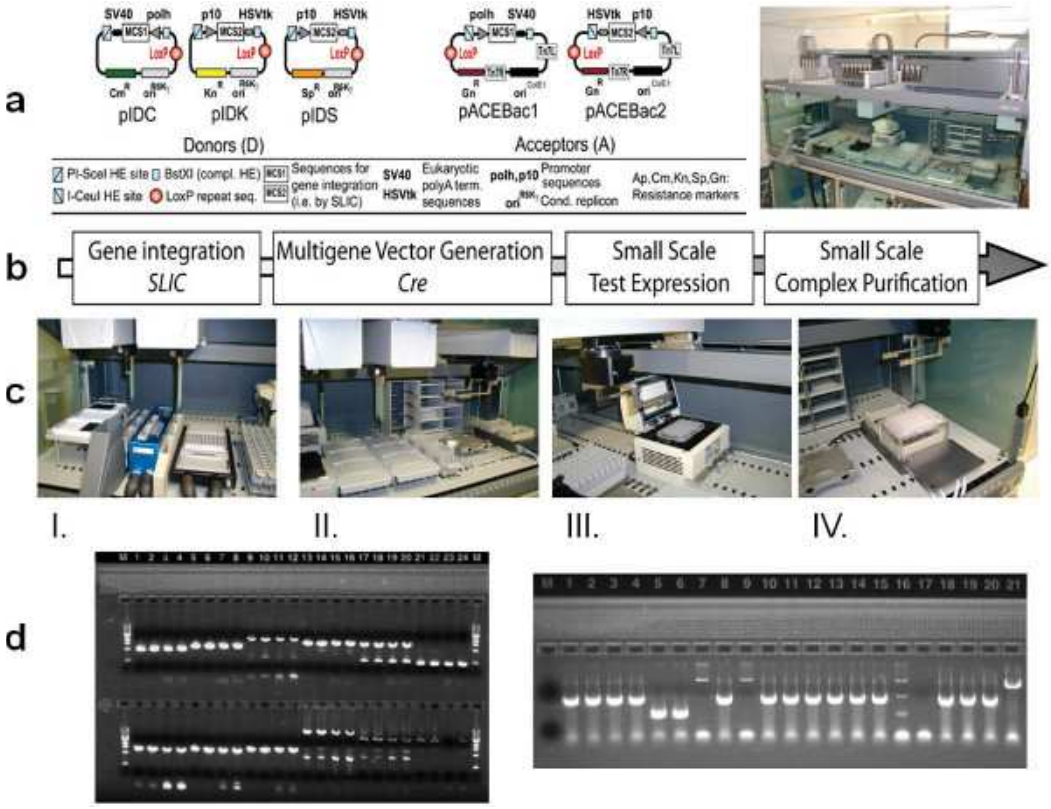


Figure 2. (Automated) Tandem Recombineering (TR). (a) Tandem recombineering utilizes small arrays of synthetic plasmid DNA molecules called Donors and Acceptors for assembling multigene expression constructs, optionally in robot-assisted mode using a liquid handling work-station (right). **(b)** The TR workflow is shown. SLIC stands for sequence and ligation independent cloning, Cre stands for the Cre-LoxP fusion concatenating Donors and Acceptors into which genes of interest have been inserted by SLIC. Multigene expression constructs generated by this recombineering procedure using SLIC and Cre in tandem are then integrated into the MultiBac baculovirus genome and used for small and large scale expression in insect cell cultures infected by the recombinant virus. **(c)** Snapshots of the robot-assisted TR process are shown including provision of template DNA and primers (I), preparation of PCR reactions in multiwell plates (II), PCR amplification of DNAs (III) and preparation of multigene constructs grown in bacterial culture by alkaline lysis in multi-well plates (IV). **(d)** PCR products used for TR are visualized by using the e-gel system (left). Completed multigene constructs are validated by analytical PCR reactions loaded on an e-gel (right). Click here to view larger figure.

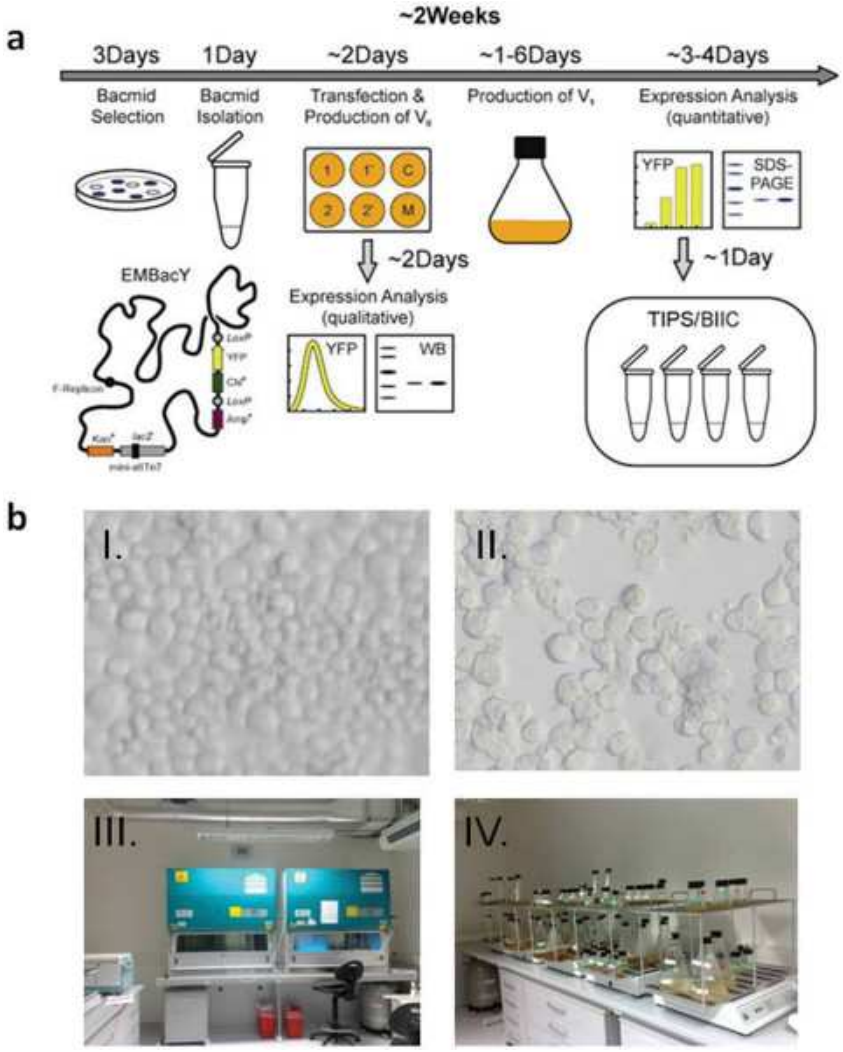


Figure 3. MultiBac virus generation, amplification, storage. (a) The standard operating procedure (SOP) of the MultiBac platform at the EMBL Grenoble is shown in a schematic view. Recombinant MultiBac virus is identified by blue-white screening and prepared from bacterial cultures. Initial transfection takes place on 6-well plates seeded with monolayers of insect cells (Sf21, Sf9, Hi5, others). Virus is amplified and target protein produced in Erlenmeyer shakers. Virus is stored by freezing aliquots of infected insect cells (BIIC). Florescence is recorded as an analytical tool if YFP (or another fluorescent protein) has been integrated as a marker protein for example into the loxP site present on the MultiBac baculoviral genome. **(b)** Snapshots of insect cells infected with MultiBac baculovirus are shown. The cells stop proliferating, increase in size (I.). Cell fusions are observed (II). Virus budded off the infected cells into the media is collected and used to infect larger cell cultures for protein complex production (III, IV). Click here to view larger figure.

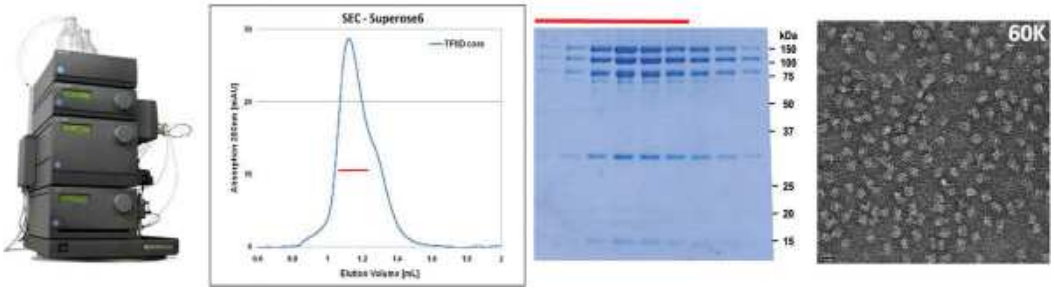


Figure 4. Protein complex production and down-stream processing. Protein complex sample can be already conveniently purified from small-scale initial cell cultures by utilizing miniaturized purification methods such as multiwall or microtip-based purification, optionally in conjunction with small-volume chromatography systems (left). Often the yield of already these "analytical" purification runs (middle) is sufficient for analyzing the structure and function of the purified complex by a variety of means including electron microscopy (right). Click here to view larger figure.

Discussion

Video snap-shots in **Figures 2** and **3** illustrate the entire process from robot-assisted generation from cDNA of multigene expression constructs all the way to infection of insect cell cultures for protein production. New reagents (plasmids and virus) and robust protocols have been developed to enable a pipeline relying on SOPs. The entire pipeline has been implemented as a platform technology at the EMBL in Grenoble. The MultiBac platform has been accessed by many scientists from academia and industry who are engaged in multiprotein research. The training access is supported by dedicated access programs funded by the European Commission (P-CUBE, BioSTRUCT-X).

The availability of SOPs to carry out protein complex expression by using the MultiBac system has rendered this technology easily amenable also to non-specialist users. Robot-assisted operation accelerates multiprotein complex production in particular when a sufficiently large number of complexes, for example variants and mutants of a specimen of interest, need to be generated in parallel. However, manual operation at low- to medium throughput also greatly benefits from the availability of SOPs. In our experience, processes that could be successfully scripted into robotics routines needed to be refined first with significant effort until sufficiently robust protocols were obtained that are compatible with using a robot. Such protocols form the basis of our SOPs.[5,6,14,15] Indeed, the implementation of these robust protocols for the robot lead to a very considerable efficiency gain also of manual operations in our laboratory.

Many proteins and protein complexes have been and are being produced by using the MultiBac system that we developed, and close to 500 laboratories world-wide have obtained the reagents. MultiBac has catalyzed research not only in structural biology but also in many other areas of life sciences that investigate or exploit the interactions between proteins in large assemblies. MultiBac has also been used to produce protein targets of considerable pharmacological interest including virus-like particles, which may become useful vaccine candidates.[4] More recently, MultiBac has also been used to deliver genes into mammalian cells and cell cultures, or even entire organisms by gene therapy.[4] We anticipate that approaches such as those illustrated in this contribution will prove to be useful for many areas of research involving multiprotein assemblies and complex interplay of biological macromolecules that form the basis of cellular processes in health and disease.

Disclosures

IB is inventor on patents and patent applications detailing parts of the technology here described.

Acknowledgements

We thank Christoph Bieniossek, Simon Trowitzsch, Daniel Fitzgerald, Yuichiro Takagi, Christiane Schaffitzel, Yvonne Hunziker, Timothy Richmond and all past and present members of the Berger laboratory for help and advice. The MultiBac platform and its development have been and are generously supported by funding agencies including the Swiss National Science Foundation (SNSF), the Agence National de Recherche (ANR) and the Centre National de Recherche Scientifique (CNRS) and the European Commission (EC) in Framework programs (FP) 6 and 7. Support for transnational access is provided by the EC FP7 projects P-CUBE (www.p-cube.eu) and BioStruct-X (www.biostruct-x.eu). The French Ministry of Science is particularly acknowledged for supporting the MultiBac platform at the EMBL through the Investissement d'Avenir project FRISBI.

References

1. Nie, Y., Viola, C., Bieniossek, C., Trowitzsch, S., Vijay-Achandran, L.S., Chaillet, M., Garzoni, F., & Berger, I. Getting a Grip on Complexes. *Curr. Genomics.* **10** (8), 558-572 (2009).
2. Robinson, C.V., Sali, A., & Baumeister, W. The molecular sociology of the cell. *Nature.* **450** (7172), 973-982 (2007).
3. Kost, T.A., Condreay, J.P., & Jarvis, D.L. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nat. Biotechnol.* **23** (5), 567-575 (2005).
4. Bieniossek, C., Imasaki, T., Takagi, Y., & Berger, I. MultiBac: expanding the research toolbox for multiprotein complexes. *Trends Biochem. Sci.* **37** (2), 49-57 (2012).
5. Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., & Berger, I. Protein complex expression by using multigene baculoviral vectors. *Nat. Methods.* **3** (12), 1021-1032 (2006).
6. Bieniossek, C., Richmond, T.J., & Berger, I. MultiBac: multigene baculovirus-based eukaryotic protein complex production. *Curr. Protoc. Protein Sci.*, Chapter 5, Unit 5.20, (2008).
7. Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., & Berger, I. New baculovirus expression tools for recombinant protein complex production. *J. Struct. Biol.* **172** (1), 45-54 (2010).
8. Vijayachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., Chaillet, M., Schaffitzel, C., Busso, D., Romier, C., Poterszman, A., Richmond, T.J., & Berger, I. Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J. Struct. Biol.* **175** (2), 198-208 (2011).
9. Thomas, M.C. & Chiang, C.M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41** (3), 105-78 (2006).
10. Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., & Ban, N. Atomic structures of the eukaryotic ribosome. *Trends Biochem. Sci.* **37** (5), 189-98 (2012).
11. Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., & Yusupov, M. One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19** (6), 560-567 (2012).
12. Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., & Kornberg, R.D. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science.* **288** (5466), 640-649 (2000).

13. Berger, I., Fitzgerald, D.J., & Richmond, T.J. Baculovirus expression system for heterologous multiprotein complexes. *Nat. Biotechnol.* **22** (12), 1583-1587 (2004).
14. Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O., & Berger, I. Automated unrestricted multigene recombineering for multiprotein complex production. *Nat. Methods.* **6** (6), 447-450 (2009).
15. Nie, Y., Bieniossek, C., Frey, D., Olieric, N., Schaffitzel, C., Steinmetz, M.O., & Berger, I. ACEMBLing multigene expression constructs by recombineering. *Nat. Protocols.*, doi:10.1038/nprot.2009.104 (2009).
16. Wasilko, D.J., Lee, S.E., Stutzman-Engwall, K.J., Reitz, B.A., Emmons, T.L., Mathis, K.J., Bienkowski, M.J., Tomasselli, A.G., & Fischer, H.D. The titerless infected-cells preservation and scale-up (TIPS) method for large-scale production of NO-sensitive human soluble guanylate cyclase (sGC) from insect cells infected with recombinant baculovirus. *Protein Expr. Purif.* **65** (2), 122-32 (2009).
17. Chao, W.C., Kulkarni, K., Zhang, Z., Kong, E.H., & Barford, D. Structure of the mitotic checkpoint complex. *Nature.* **484** (7393), 208-13 (2012).
18. Yamada, K., Frouws, T.D., Angst, B., Fitzgerald, D.J., DeLuca, C., Schimmele, K., Sargent, D.F., & Richmond, T.J. Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature.* **472** (7344), 448-53 (2011).

PUBLICATION 4

Gene gymnastics: Synthetic biology for baculovirus expression vector system engineering.

Authors: *Lakshmi Sumitra Vijayachandran, Deepak Balaji Thimiri Govinda Raj, Evelina Edelweiss,* ***Kapil Gupta****, Josef Maier, Valentin Gordeliy, Daniel J Fitzgerald and Imre Berger*

Bioengineered; Volume 4, Issue 5, September/October 2013, Pages 279-287
DOI: 10.4161/bioe.22966

The following manuscript describes a developed Baculovirus Expression Vector System (BEVS) - OmniBac, for better recombinant protein expression. This article also explains an approach for generating synthetic baculovirus genome for improved protein expression.

Bioengineered 4:5, 279–287; September/October 2013; © 2013 Landes Bioscience

Gene gymnastics

Synthetic biology for baculovirus expression vector system engineering

Lakshmi S Vijayachandran,[1,†] Deepak B Thimiri Govinda Raj,[1,†] Evelina Edelweiss,[1,2] Kapil Gupta,[1] Josef Maier,[3] Valentin Gordeliy,[2] Daniel J Fitzgerald[4] and Imre Berger[1,*]

[1]European Molecular Biology Laboratory (EMBL); Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI); UJF-EMBL-CNRS, UMR 5233; Grenoble, France; [2]Institut de Biologie Structurale (IBS); UMR5075 CEA-CNRS-Université Joseph Fourier; Grenoble, France; [3]Information Services to Life Science (IStLS); Oberndorf am Neckar, Germany; [4]Geneva Biotech; Geneva, Switzerland

[†]These authors contributed equally to this work.

Most essential activities in eukaryotic cells are catalyzed by large multiprotein assemblies containing up to ten or more interlocking subunits. The vast majority of these protein complexes are not easily accessible for high resolution studies aimed at unlocking their mechanisms, due to their low cellular abundance and high heterogeneity. Recombinant overproduction can resolve this bottleneck and baculovirus expression vector systems (BEVS) have emerged as particularly powerful tools for the provision of eukaryotic multiprotein complexes in high quality and quantity. Recently, synthetic biology approaches have begun to make their mark in improving existing BEVS reagents by de novo design of streamlined transfer plasmids and by engineering the baculovirus genome. Here we present OmniBac, comprising new custom designed reagents that further facilitate the integration of heterologous genes into the baculovirus genome for multiprotein expression. Based on comparative genome analysis and data mining, we herein present a blueprint to custom design and engineer the entire baculovirus genome for optimized production properties using a bottom-up synthetic biology approach.

Keywords: baculovirus, BEVS, multiprotein complexes, MultiBac, OmniBac, comparative genome analysis, recombinant protein production

Submitted: 10/26/12

Revised: 11/19/12

Accepted: 11/19/12

http://dx.doi.org/10.4161/bioe.22966

*Correspondence to: Imre Berger; Email: iberger@embl.fr

Our understanding of the cellular machinery has increased tremendously in recent years mainly due to astounding progress in "omics" research (genomics, proteomics and glycomics).[1] Comprehensive genomics data sets are now available for many organisms including human, and focus has now shifted to elucidating the cellular proteome in correlation with the live cellular functionality and morphology. One essential lesson already learned is that proteins in eukaryotic cells typically do not work in isolation but coexist in large and highly diverse assemblies of ten or more interlocking subunits. These stably or transiently associated multiprotein assemblies additionally work together with separate proteins or multiprotein assemblies to carry out essential cellular processes including signaling, energy generation, and transport of food, water or waste. The highly dynamic processes going on among the proteins in our cells has been termed the "protein sociology."[2]

A considerable number of accessory proteins typically accompany any individual multiprotein complex at various stages of its production, trafficking, active life and degradation. For example, chaperones are often critical for proper assembly of complexes, while other proteins are required for proper targeting and activation through post-translational modification. The activity of complexes is often fine-tuned by the incorporation of isoforms of individual subunits, for example to mediate tissue-specific functions. To fully understand biology, it is clear that we need methods to unlock the assembly, structure and mechanism of all of the complexes that exist in our cells. This is

not only essential for basic research, but equally important for enabling novel approaches in the pharmaceutical and biotech industries to drive development of new and better drugs that more specifically modulate cellular functions.[3] An imposing bottleneck that obstructs progress in these areas stems from the typically low abundance and high heterogeneity of protein complexes in their native cells. Apart from a handful of notable examples, most human multiprotein complexes remain virtually inaccessible to date.

Recombinant overexpression can provide a solution to this problem. However, until recently, the production challenge for eukaryotic (especially human) multiprotein complexes has not been systematically addressed by the community. There is no denying that especially for drug design, it is imperative that we no longer recoil from studying human protein complexes, given that most surrogate systems such as the fruit fly, yeast or archea which are commonly used in basic research do not accurately recapitulate the functions of the much higher developed and therefore significantly more complicated human species.

The provision of human multiprotein complexes in the quality and quantity required for mechanistic studies and drug design poses particular challenges due to the complexity of the machinery at work in our cells. Technical factors for heterologous protein production including protein yield, stoichiometric ratio between subunits, post-translational modifications, folding, and stability are all of critical importance, and ideally a highly flexible heterologous expression system should be available that can provide these functions for a wide range of protein complexes. An attractive solution could be mammalian expression systems, which naturally provide the required functions to accurately reflect what takes place in our cells, and heterologous expression in mammalian systems has become increasingly popular, especially for secreted proteins such as therapeutic antibodies.[4] However, mammalian systems often do not provide acceptable yields for intracellular proteins, and multiprotein expression technologies for mammalian cells are still in their infancy, albeit progress has been made recently, opening interesting options to depict with hitherto unattainable precision entire pathways in mammalian cells, for example for pharmacological screening studies.[5]

An attractive alternative to mammalian systems is heterologous expression using recombinant baculoviruses to infect insect cell cultures. This method was pioneered three decades ago[6] and has become a method of choice for producing high levels of many eukaryotic proteins including a large number of proteins of pharmaceutical interest.[7,8] A significant advance over existing baculovirus expression vector systems (BEVS) came with the introduction of MultiBac, an advanced BEVS particularly tailored for producing eukaryotic multiprotein complexes for structural and functional studies.[9-12] MultiBac consists of a baculovirus genome that has been engineered for optimized protein production by deleting protease and apoptosis activities.[13] In a subsequent improvement of the system, a new suite of transfer vectors was introduced to facilitate introduction of many heterologous genes into one recombinant MultiBac baculovirus by a method called tandem recombineering (TR), involving sequence-and-ligation-independent cloning (SLIC) and Cre-LoxP recombination.[14,15] More recently, the design of transfer plasmids has been further refined, resulting in small, easy to handle plasmids containing only the functional DNA elements required for protein expression, expression cassette multiplication and plasmid concatenation by TR.[15] Multigene transfer vectors created in this way are introduced into the Multi Bacbaculovirus genome by the Tn7 transposon, in *E. coli* strains modified for this purpose.[14]

As a step forward relative to previous systems, the original MultiBac system already provided the option to integrate accessory functionalities that may be required for proper functioning of a multiprotein complex, by means of a second entry site engineered into the virus genome that is independent of, and distal to the main site of integration that relies on the Tn7 transposition. This feature has been exploited to integrate additional functional modules into the viral genome, including post translational modification enzymes, and fluorescent proteins that allow easy monitoring of virus performance and protein production following transfection and during virus amplification.[15,16] More recently, this approach has been used to create SweetBac, allowing for the production of mammalian-like glycoproteins in insect cells.[17,18] MultiBac is now in use at more than 600 laboratories world-wide, in academia and industry, and a broad range of multiprotein complexes have been produced in high quality and quantity for diverse applications by using the MultiBac system.[3,9]

Currently, two approaches for integrating heterologous expression cassettes into the baculovirus genome dominate the field. One of these approaches requires the presence of the baculoviral genome as a bacterial artificial chromosome (BAC) in *E. coli* cells, together with Tn7 transposase activity present in these same cells which recombine transformed transfer plasmids into a Tn7 attachment site on the BAC. Invitrogen's Bac-to-Bac system and also the more advanced MultiBac system both utilize this approach. The recombinant, composite baculovirus DNA is then purified from these *E. coli* cells by alkaline lysis, and used to transfect insect cells. In contrast, the original method of choice to integrate heterologous expression cassettes into the baculovirus genome relied on homologous recombination mediated by regions in the transfer plasmid that were homologous to two genes on the baculovirus genome (Orf1629 and lef2/603) that flank the baculoviral polh locus which had been inactivated. This method is still offered by a large number of commercial providers (Novagen BacVector series, PharMingen BaculoGold, Abvector and others). By this method, homologous recombination occurs in insect cells following transfection the baculovirus genomic DNA together with the transfer vector. The efficiency of recombination is increased by linearization of the baculovirus genome, but still remains a less efficient method to rapidly generate recombinant baculovirus than transforming Tn7-produced composite BACs. A further improvement on the homologous recombination in insect cell method came by truncation of the essential Orf1629 gene on the baculovirus genome which is then repaired by co-transfecting complete Orf1629-containing

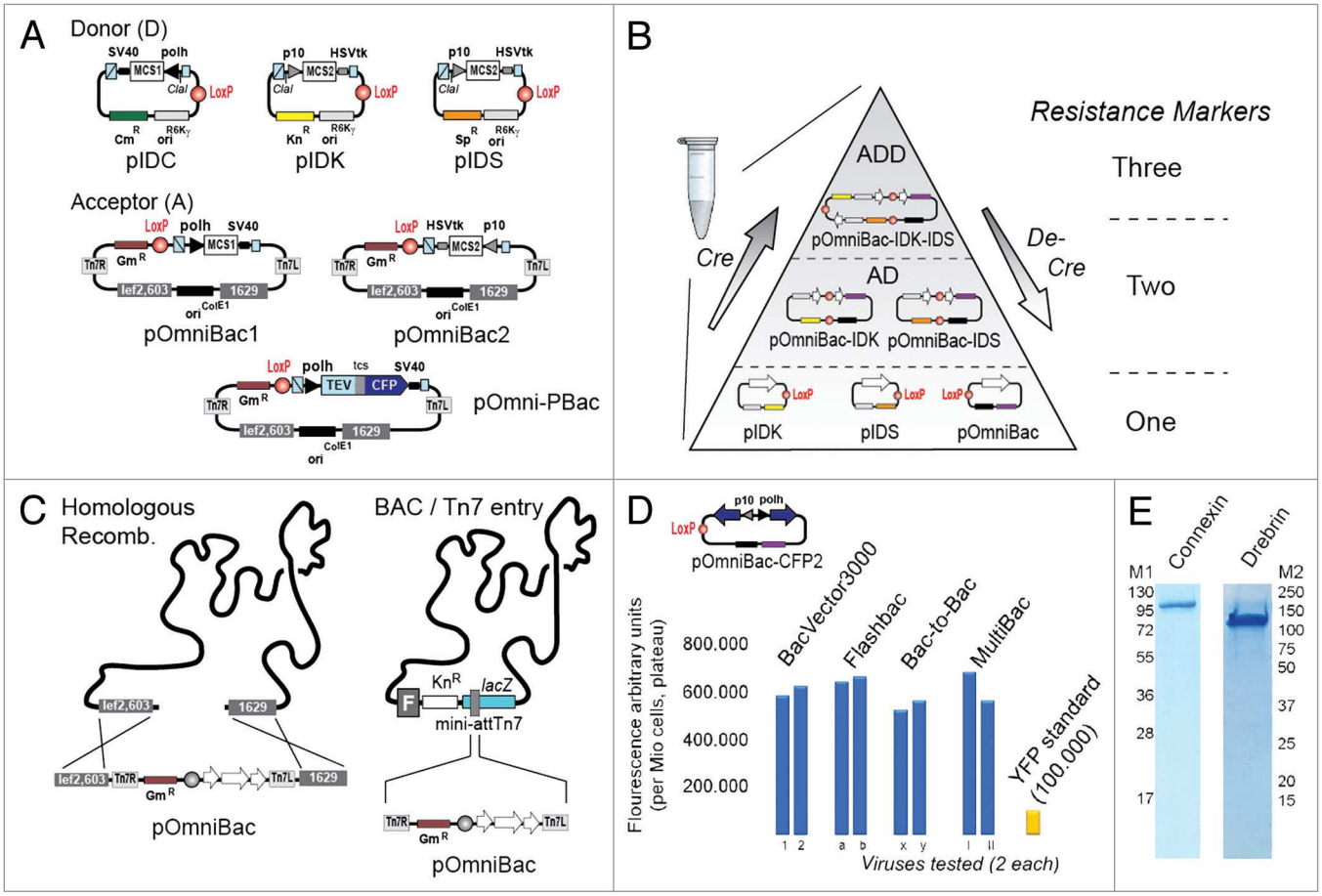


Figure 1. OmniBac. (**A**) Synthetic donor and acceptor plasmids are shown schematically. DonorspI DC, pIDK and pIDS are from the original MultiBac suite.[9] Novel acceptor plasmids pOmniBac1, pOmniBac2 and pOmni-PBac are shown. Cm stands for chloramphenicol, Kn for kanamycin, Sp for spectinomycin, Gm for gentamycin. Viral very late promoters polh and p10, and SV40 and HSVtk polyadenylation signals are indicated. Donors contain a conditional origin of replication derived from the R6Kγ phage.[9] A multiplication module comprising a homing endonuclease site and a complementary BstXI site is shown as boxes in light blue, flanking the expression cassettes.[14] MCS1 and MCS2 denote multiple cloning sites.[14,15] Acceptors contain a regular ColE1 origin of replication, the DNA sequence elements for Tn7 transposition (Tn7L, Tn7R) and in addition the lef2/603 and Ori1629 homology regions (gray boxes). All plasmids have a LoxP sequence for plasmid fusion mediated by Crerecombinase. pOmni-PBac contains the polyprotein expression cassette for multiprotein complex expression including a gene encoding for tobacco etch virus Nia protease (TEV) followed by a fluorescent marker, cyan fluorescent protein (CFP), spaced by a TEV protease cleavage site (tcs). Polyprotein constructs are generated in pOmni-PBac as described previously for our pPBac design.[15] (**B**) OmniBac acceptors and donors are concatenated by Cre-LoxP fusion in a combinatorial fashion.[14,15] For matter of clarity, only two donors are shown reacting with one acceptor, each containing one heterolgous gene (white arrow). Unfused donors act as suicide plasmids upon transformation due to their conditional origin.[15] AD and ADD fusions, in contrast, are efficiently propagated based on the regular origin in the acceptor used.[15] Plasmid combinations are selected based on unique resistance marker combinations.[15] (**C**) OmniBac acceptors contain all elements required for generating recombinant baculoviruses for expression by using the homologous recombination method (left) or, alternatively, the BAC/Tn7 entry method (right). (**D**) An OmniBac plasmid containing two genes encoding for CFP (top) was used to test homologous recombination entry (BacVector3000, FlashBac) as well as the BAC/Tn7 entry (Bac-to-Bac, MultiBac). Quantification of fluorescence on an arbitrary scale calibrated to a yellow fluorescent protein standard (YFP) and 1 million infected cells each exhibited virtually identical heterologous expression levels (two clones each were tested). (**E**) Connexin (left) and Drebrin (right) were expressed using OmniBac and purified to homogeneity. Connexin migrates as an oligomer in SDS-PAGE.

transfer vectors (FlashBac system, Oxford Expression Technologies).

OmniBac: Multigene Transfer Plasmids to Access All Available Baculoviruses

Currently, available baculovirus transfer vectors are compatible with either either baculovirus genomes using Tn7 for gene entry or baculovirus genomes using homologous recombination, respectively. However, no transfer vectors can access both types of genomes. We perceived this situation as unsatisfactory, given that both systems (BAC/Tn7 entry and homologous recombination, respectively) offer unique beneficial opportunities and particular baculovirus genomes offer specific advantages in production of different classes of recombinant proteins. The BAC/Tn7 entry system, for instance, is thought to be more accessible to non-specialist users without experience in cell culture techniques. The homologous recombination system, on the other hand, has the advantage that it can be efficiently scripted in

high-throughput robotic routines, which is basically unfeasible with the BAC/Tn7 entry system.[19] Automation, however, is indispensable to accelerate modern protein research.

We have resolved this issue by creating new "OmniBac" transfer plasmids (**Fig. 1**, **Supplemental Materials**). The OmniBac vector suite contains new custom designed acceptor plasmids, which provide both the Tn7 transposon sequences (Tn7L and Tn7R), and at the same time also contain the Orf1629 and lef2/603 sequences required for homologous recombination (**Fig. 1A**). The basic design of these Acceptor plasmids follows the logic of the original MultiBac acceptor plasmids[9] and thereby facilitates multiprotein expression. The newly synthesized pOmniBac1 and pOmniBac2 plasmids contain all elements required for gene insertion by SLIC, the homing endonuclease/BstXI based module to facilitate multiplication of expression cassettes on each transfer plasmid and the LoxP sequence for Cre catalyzed fusion of Donor plasmids (pIDC, pIDK and pIDS) with TR offering the option to add further genes (**Fig. 1B**). The donor plasmids are from the recent MultiBac Acceptor/Donor suite.[15] OmniBac Acceptor based multigene transfer plasmids can access all currently available baculoviruses using either homologous recombination, or the BAC/Tn7 entry approach (**Fig. 1C**).

In a proof of concept, we integrated two CFP genes, one under control of the viral very late promoter polh, and one under control of the viral very late promoter p10 by SLIC, giving rise to pOmniBac-CFP2 (**Fig. 1D**). We reacted the pOmniBac-CFP2 plasmid with several different baculoviruses using the Tn7 entry method (Bac-to-Bac, our MultiBac) or the homologous recombination method (BacVector3000, FlashBac), and monitored protein production from our OmniBac expression vectors by CFP flourescence. We used YFP which is integrated into our MultiBac baculoviral genome as a marker for virus performance and protein expression (intensity arbitrarily set to 100,000). All viruses performed excellently in this comparison, and virtually identical levels of CFP production were reached regardless of the virus used in this experiment (**Fig. 1D**). We also scrutinized the OmniBac plasmids for Cre-LoxP recombination and production of a test membrane protein (connexin). Connexins are vital proteins mediating cell-to-cell communication.[20] Connexin could be purified to homogeneity following expression from a pOmniBac1-Connexin transfer plasmid transposed into the MultiBac virus (**Fig. 1E**). Drebrin is a protein which is thought to interact with connexins and actin.[21] Cre-LoxP fusion of a donor plasmid containing Drebrin (pIDC-Drebrin) with pOmniBac1 and subsequent Tn7 transposition into the MultiBac virus yielded a composite virus with excellent protein production characteristics (**Fig. 1E**). In summary, the OmniBac acceptor plasmids are functional for protein expression and for acceptor–donor recombineering, and can efficiently substitute for the previous MultiBac acceptor plasmids (pACEBac1 and pACEBac2) for multigene construct generation by TR. As we show here, the resulting multigene transfer constructs can, for the first time, access all baculoviruses that are presently available. The OmniBac suite thus significantly expands the scope of possible multiprotein production experiments using BEVS, now providing the option to choose from among all available baculovirus genomes the one most suitable for the specific experiment at hand.

Baculovirus Genome Engineering

Currently, BEVS applications including MultiBac rely on a large baculovirus genome (~130 kb) derived from wild-type *Autographa californica* multicapsid nuclear polyhedrosis virus (AcMNPV). This genome has been intensively researched for many years. Genes that are essential for propagation in cell culture and genes which are detrimental for foreign protein production were delineated by several research groups.[22-28] Based on our positive experiences with bottom-up design of plasmids and engineering of the baculovirus genome itself to improve protein production, we became interested in extending our reengineering concept to rewiring the entire baculovirus genome to maximize its performance. Redesigning and restructuring the baculovirus genome for enhanced DNA stability and efficient protein production requires deletion, modification or insertion of DNA sequences. So far, the existing alterations of the wild-type genome were made by classical knockout strategies, and were mostly directed at changes to the genome to facilitate gene insertion, and to improve protein production by removing detrimental proteins such as virally encoded protease and chitinase. There is a large landscape for engineering the baculovirus genome, given the many non-essential genes that are probably dispensable for virus maintenance and heterologous protein expression. Also, the inherent DNA instability of the current baculovirus genome poses a problem, in particular at expression scales relevant for pharmaceutical production. Simply speaking, as the virus replicates during expression scale up, it progressively suffers from deletion of bits and pieces of its genome, preferentially in the highly expressed, (non-essential) heterologous protein expression cassette, as we have shown already for laboratory scale production.[12,23,24] This is exacerbated for the viruses of the BAC/Tn7 type by the fact that the insertion site which is targeted by the Tn7 transposon is actually a mutational hotspot.[29,30] It is clear that the possibilities to improve the baculovirus itself are ample, and are far from exhausted in state of the art systems.

Analogous to our approach to minimize transfer plasmids which then could be synthesized de novo, we became interested in the genes and DNA elements which are dispensable under laboratory culture conditions and unnecessary for efficient budded viral production, which is the major virus type used for protein expression in cell culture.[11,12] We reasoned that conceivably, the baculovirus can be engineered by removing non-essential genes and regions prone to mutation, with possibly large benefits for virus DNA stability, and also for accommodating very large foreign gene insertions, without compromising the ease of handling and the superior protein production properties offered by the system. For illustration, the polyhedrin gene locus (polh) is a non-essential viral gene that codes for the polyhedrin protein which produces occlusion bodies. These occlusion bodies can be

observed as large particles in the insect cell nuclei at the very late stage of the infectious cycle. The discovery of polyhedrin as a nonessential gene propelled the generation of BEVS.[31,32] The majority of heterologous proteins produced by BEVS utilize a baculoviral genome in which the polh gene is deleted and foreign genes of interest are produced under the control of the polh promoter. Currently, there are several BEVS with other non-essential gene deletions such as chitinase (v-cath), cathepsin (chiA), p10, p74 and p26 (FlashBacUltra, OET; BacVector3000, Novagen). If the BEVS genome were reengineered, or even designed entirely de novo from scratch, all nonessential genes and known mutational hot spots could in theory be deleted, and all existing entry sites and beneficial modifications integrated into a new baculovirus genome with superior characteristics to anything yet conceived or developed.

A Blueprint for a Synthetic Baculovirus Genome

We performed extensive literature mining on the AcMNPV genome in identifying essential and non-essential genes with respect to cell culture propagation, budded virus production and foreign protein expression. We included comparative analysis on neighboring genomes, which belong to the so-called clade I/Ib alpha baculovirus group encompassing BmNPV, BmaNPV, MvMNPV, PlxyMNPV and RoMNPV. Besides the comparative genome studies, other analyses scrutinized gene synteny, promoter motifs, repetitive features in origins of replication that are involved in homologous recombination events, and potential transposon integration sites. These were then all compiled and integrated into our baculovirus genome database. Based on these analyses and on Genebank nucleotide RefSeq entry NC_001623.1, we compiled an annotated baculovirus genome map as a blueprint for future genomic engineering (**Fig. 2**).

In this baculovirus genome map, we show in total 156 (ORF) genes (**Fig. 2**). Out of these, 62 genes could be classified as (likely) non-essential genes which are probably dispensable (annotated by shades of red color). Ninety-four genes, in contrast, were classified as essential genes (annotated by shades of green color to black) for cell culture propagation. The essential genes were classified based on conservation which includes core genes, core genes with fixed synteny (co-localization of loci within the species), conserved or variable (no-deletion category in clade-I/II alpha baculovirus group), and unique in AcMNPV. Additionally, 9 homologous regions (hr) are shown in the map, which serve as transcriptional enhancers and origins of replication, some of which may not be essential. Genes were classified as non-essential based on the function in oral infectivity, host cell interactions, cell lysis and moulting inhibition. Non-essential genes also include genes that are involved in inducing apoptosis such as v-cath, genes and genes that were previously reported not to give rise to a deletion phenotype.

Interestingly, when we analyzed the distribution of these annotated essential and non-essential genes, we found that a majority of the essential genes was localized in one half of the baculovirus genome. In contrast, the majority of the genes that could likely be disposed of are found clustered in the other half of the genome. Based on this distribution of essential and non-essential genes, we divided the circular baculovirus genome map into two almost equal sized semicircles (**Fig. 2**). We observed that close to 55 (35%) of essential genes and 22 (14%) of non-essential genes were located in one semicircle (upper semicircle, **Fig. 2**), while 35 (22%) of non-essential genes and 44 (28%) of essential genes were located in the other semicircle (lower semicircle, **Fig. 2**). In addition, the majority of the currently existing BEVS have gene loci that are either deleted or modified in the upper semicircle (i.e., MultiBac, v-cath and chiA deleted; MultiBac and Bac-to-Bac, Tn7 attachment site replaces polyhedrin site; FlashBac, ORF1629 and polyhedrin site modified; FlashBacULTRA, p10, P74 and p26 genes deleted). This interesting asymmetric distribution of essential and non-essential genes provides ample opportunities to engineer the baculoviral genome by bringing an entire array of synthetic biology methods to bear, to design and synthesize a minimal and functional genome, i.e., a baculovirus genome containing only the essential (coding and non-coding) genetic elements required for its life cycle under laboratory culture conditions and supporting the highest possible expression levels of functional proteins and protein complexes.

Synthetic Biology: Toward an Optimized Baculovirus Genome

Synthetic biology creates artificial agents that recapitulate the behavior of living systems including inheritance, genetics and evolution.[33] Breakthroughs for synthetic biology include novel synthetic genomes with promising phenotypes such as engineered bacteriophages to combat antibiotic-resistant bacteria, inhibition of certain bacterial genetic programs that could improve the efficacy of antibiotic therapies[34] and viruses that were customized to invade cancer cells.[35] Craig Venter and his associates have reported the design and synthesis of a 1 Mb *Mycoplasma mycoides* bacterial genome and also of an entire synthetic bacteriophage genome.[36,37] Similar approaches can be utilized also for the synthesis and assembly of an optimized, synthetic baculovirus genome. The baculovirus genome, with ~130 kb size, is small compared with the genome of *Mycoplasma*. By using the methods pioneered by the Venter group, wild-type segments in the baculovirus genome could be iteratively replaced with custom designed synthetic DNA fragments, using for example *E. coli* or yeast based recombination as a method of choice for this.[38]

An alternative, potentially promising approach would be to use evolution by natural selection to generate "minimal and functional baculovirus genomes". It has been shown that baculoviruses, when serially passaged in cell culture, are prone to massive auto deletions, resulting in virus populations that are dominated by non-functional, small circular entities that rely on the presence of a few functional viruses that supply the proteins required for replication. It will be interesting to see whether or not evolution experiments can be designed which lead to intermediate deletion genomes that are significantly reduced in size as compared with the wild-type baculovirus, but are still infectious and able to propagate

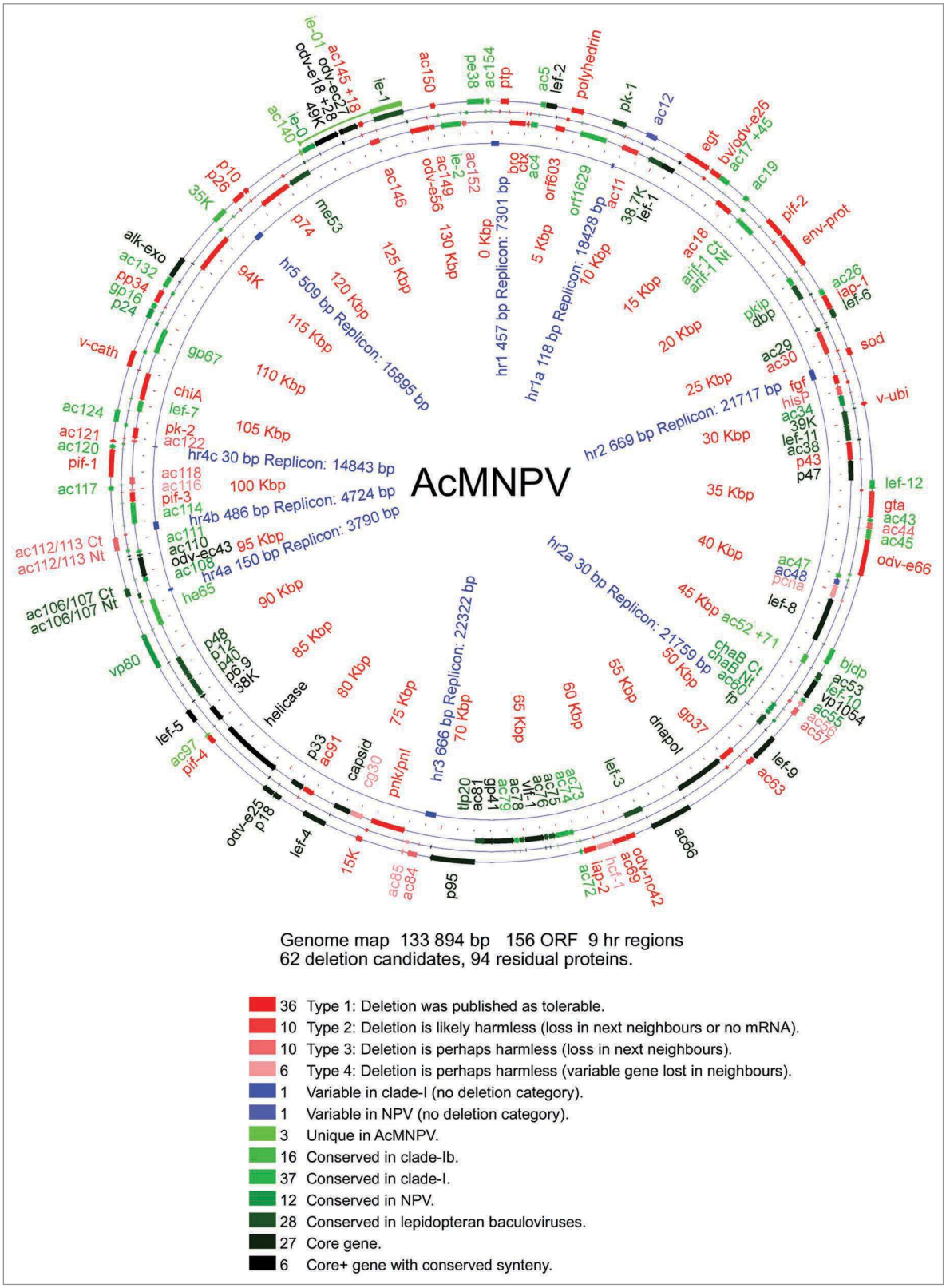


Figure 2. Blueprint of the baculovirus genome. The annotated genome of *Autographa californica* multicapsid nuclear polyhedrosis virus (*Ac*MNPV) is shown in a schematic representation (top). Genes are scored (bottom) on a scale from essential genes that are conserved (no deletion category, shades of green color to black) to non-essential genes (shown to be possibly deleteable, shades of red color). The classification applied is detailed in the legend (bottom). The annotated genome map was generated by a self-developed Perl program. Essential and non-essential genes are not randomly distributed but cluster in the genome. The upper semicircle is composed to 56% of the essential genes (and 44% of non-essential genes), the lower, more conserved semicircle is composed to 71% of essential genes (and 29% of non-essential genes).

and produce heterologous protein in cell culture. Alternatively, specimens may be isolated that only need to be reengineered by adding a few genes to restore the desired functions. This could possibly be achieved by serial passaging of baculovirus in insect cells and investigating the distribution of genome sizes within a heterogeneous population of serially passaged viruses.[39] Evidence exists that over time baculoviruses have the tendency to restructure their genome by discarding genes by natural recombination events,[23] resulting in defective-interfering particles and other minimal baculovirus genome that are not well characterized. It will be potentially highly beneficial to revisit such virus subpopulations and compare their phenotype with wild-type in terms of stability, functionality and quality of budded virus, thereby identifying additional features that may improve heterologous protein production.

Future Perspective

BEVS is emerging increasingly as a method of choice for the production of eukaryotic proteins for a wide range of applications in academic and industrial research and development. Since its inception more than three decades ago, BEVS have been improved, one step at a time, for simplifying handling and optimizing protein production properties. More recently, the introduction of the MultiBac system has contributed to unlocking the realm of large multiprotein complexes to functional studies to understand protein sociology and biological function in the cell, and to develop new and better drugs by exploiting, in the future, the tremendous amount of data amassed through the genomics and proteomics studies that increasingly dominate current research efforts in the life sciences.

BEVS until today came mostly in two flavors, differing in their heterologous gene integration strategies (BAC/Tn7 entry and homologous recombination, respectively) that were mutually exclusive. As a consequence, the initial transfer plasmid into which the genes of choice were inserted predetermined which baculovirus genomes were available for the protein expression experiment. A number of virus variants exist under each of the two categories, each with its own merits. We have presented here our novel OmniBac plasmids, which combine the DNA elements required for both gene integrations strategies, and are therefore compatible with all currently available baculoviruses. The OmniBac plasmids were shown to function equally well with all baculoviruses tested. Further, they are also fully compatible with the tandem recombineering approach we developed to facilitate multigene insertion into the baculovirus of choice. We anticipate that the OmniBac plasmids will in the future substitute for currently used transfer plasmids for many applications, including high-throughput pipelines that are becoming increasingly utilized for heterologous expression of eukaryotic target proteins in insect cells.

We hypothesize that the custom design strategies that lead us to minimal transfer plasmids such as OmniBac and the other components of the MultiBac transfer plasmid suite will also be beneficial if applied to the baculovirus genome itself, which is currently only marginally optimized for heterologous protein production. Toward this end, we have used comparative genomics and data mining, and realized that the essential genes on the one hand, and the non-essential genes on the other, are clustered together on the baculovirus genome. This opens up interesting possibilities to rewire the genome, potentially also entirely by design-from-scratch in a bottom-up approach, by applying synthetic biology techniques for genome engineering. This endeavor will be complemented by in vitro evolution, using serial passaging and exploiting the innate tendency of the baculovirus to rearrange and delete segments of the genome in the process. Together, these two approaches should prove to be effective in creating a new and synthetic, minimal baculovirus which can then be customized for a variety of applications, for example by introducing modifiers and accessory proteins into the viral backbone that assist assembling an active and functional multiprotein complex of choice.

Improving and optimizing recombinant protein expression using BEVS is not limited by modifying solely the baculovirus genome itself, but can likewise include adapting the host system (insect cells). For example, insect cell lines expressing enzymes that reconstitute the human glycosylation pathway, were generated by genomic engineering.[8,40] With BEVS becoming increasingly prominent in recent years, we believe there are ample opportunities and a substantial scope for new and improved, customized baculovirus genomes, genetically modified cell lines and optimized laboratory culture procedures tailored for high performance heterologous protein production, to accelerate academic basic research and likewise protein-based drug discovery in the pharmaceutical and biotechnology industries in the near future.

Materials and Methods

OmniBac plasmids were custom designed from scratch and assembled from synthetic DNA fragments (Genscript Corp.) using standard DNA procedures. Plasmid maps and sequences are provided in the Supplement. Generation of composite baculoviruses (MultiBac, Bac-to-Bac, FlashBac, BacVector3000) was performed using published protocols[12] or following the manufacturers' recommendation. CFP fluorescence was recorded as described previously.[13] Full-length Connexin 36 (GenBank Nr. BC069339) was cloned into pOmniBac1 via EcoRI and HindIII, and a calmodulin binding peptide (CBP) and deca-histididine tandem affinity purification tag was the inserted via DraI and ScaI to yield construct pOmniBac-Connexin. Full-length Drebrin E (GenBank Nr. D17530.1) with an N-terminal strepatividin binding peptide (Strep) tag spaced by a TEV site was cloned into pIDC via XhoI and KpnI giving rise to construct pIDC-Drebrin. Gene insertions into the MultiBac viral genome, virus generation, amplification and protein expression were performed as described.[11] For purification of Connexin36 and preparation of gap junctions, Sf21 cells expressing the protein were re-suspended in bicarbonate buffer (1 mM sodium bicarbonate, pH 8.0, 8.7% sucrose) and lysed by sonication for ten seconds. Gap junctions were purified by sucrose gradient centrifugation at 192,000 g overnight, dialysed against

sodium bicarbonate solution to remove sucrose and solubilised overnight with 2% n-Dodecyl β-D-maltoside (DDM) and further purified by metal affinity purification (Ni-NTA, Qiagen). Drebrin was purified by using Strep-Tactin affinity resin (IBA GmbH) followed by size exclusion chromatography (SEC) using a Superose 6 column. Proteins were analyzed by 12% SDS-PAGE.

The data used for creating the baculovirus genome map shown in **Figure 2** were derived from mining 1253 relevant papers found in NCBI PubMed and published up to October 2011 on information for, e.g., genome sequences, gene essentiality and conservation, protein product function and localization, protein-protein interactions, mRNA expression and gene regulation. All available database annotations on gene product function were collected from NCBI Genebank, NCBI Protein, NCBI VOG clusters of related viral proteins and the UniProt database by Perl programs. Additionally 53 baculovirus genomes available in October 2011 in the NCBI RefSeq nucleotide database were analyzed for orthologous protein genes by clustering protein sequences downloaded from the NCBI protein database with a Perl program, which piped the clustering program blastclust, a part of the NCBI C-toolkit legacy blast package. Categories for conservation and variability were assigned for following different lineages of baculoviruses: all baculoviruses with conserved synteny (core+), all baculoviruses (core), lepidopteran baculoviruses, NPV (alphabaculoviruses), NPV clade I, NPV clade Ib, variable for next neighbors of AcMNPV, unique for AcMNPV. For gene essentiality classification the grades of gene conservation were integrated with published information on mRNA expression (no mRNA observed), expression in different developmental stages (immediate early, early, late, very late), content of stage-specific promoter motifs in the upstream regions and gene product function and localization indicating non-essentiality for virus propagation in cell culture (e.g., host interaction factors, oral infectivity factor, occlusion-derived virus or occlusion-body localized, other auxilary proteins, some genes acquired from the host genome). Protein genes, which were already published as non-essential, were categorized as type 1 deletion category (deletion is harmless), that having no known mRNA expression or have been proven as non-essential in the next relatives of AcMNPV (like BmNPV) as type 2 (deletion is likely harmless), that having presumably non-essential functions and are variable in alphabaculoviruses or such protein genes having unknown functions and are variable in next neighbors as type 3 (deletion perhaps harmless) and that with suspect non-essential functions and variability in other alpha baculovirus genomes but conservation in next relatives as type 4 (deletion perhaps harmless). The genome map was drawn by designing a Perl program, which imports a table of the categorized gene data and exports an image made by the Perl packages GD and GD::Simple.

Further information can be obtained upon request (iberger@embl.fr).

Disclosure of Potential Conflicts of Interest

IB and DJF are authors on patents and patent applications describing parts and applications of the technologies in the present manuscript.

Acknowledgments

We thank all members of the Berger laboratory for helpful discussions. This work was supported by the Centre National de Recherche Scientifique (CNRS) through a PEPS discovery grant (to IB), and by the European Commission (EC) Framework Program 7 (PCUBE, BioSTRUCT-X and ComplexINC, to IB). DBTGR is recipient of an EC/EMBL CoFund EIPOD fellowship.

Author Contributions

IB conceived the research with input from DJF and DBTGR; LSVA created and validated the OmniBac plasmids; DBTGR and JM performed and integrated genome analysis and data mining; JM created annotated baculovirus genome map; EE and VG used OmniBac to produce connexins and drebrin; DBTGR created pOmniBac-PBac; DBTGR, JM and IB interpreted and analyzed data; DBTGR, DJF and IB wrote the manuscript together.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/bioe/article/22966

References

1. Nie Y, Viola C, Bieniossek C, Trowitzsch S, Vijay-Achandran LS, Chaillet M, et al. Getting a grip on complexes. Curr Genomics 2009; 10:558-72; PMID:20514218; http://dx.doi.org/10.2174/138920209789503923.
2. Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007; 450:973-82; PMID:18075576; http://dx.doi.org/10.1038/nature06523.
3. Trowitzsch S, Palmberger D, Fitzgerald D, Takagi Y, Berger I. MultiBac complexomics. Expert Rev Proteomics 2012; 9:363-73; PMID:22967074; http://dx.doi.org/10.1586/epr.12.32.
4. Nettleship JE, Assenberg R, Diprose JM, Rahman-Huq N, Owens RJ. Recent advances in the production of proteins in insect and mammalian cells for structural biology. J Struct Biol 2010; 172:55-65; PMID:20153433; http://dx.doi.org/10.1016/j.jsb.2010.02.006.
5. Kriz A, Schmid K, Baumgartner N, Ziegler U, Berger I, Ballmer-Hofer K, et al. A plasmid-based multigene expression system for mammalian cells. Nat Commun 2010; 1:120; PMID:21081918; http://dx.doi.org/10.1038/ncomms1120.
6. Summers MD. Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides. Adv Virus Res 2006; 68:3-73; PMID:16997008; http://dx.doi.org/10.1016/S0065-3527(06)68001-9.
7. Kost TA, Condreay JP, Jarvis DL. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat Biotechnol 2005; 23:567-75; PMID:15877075; http://dx.doi.org/10.1038/nbt1095.
8. Jarvis DL. Baculovirus-insect cell expression systems. Methods Enzymol 2009; 463:191-222; PMID:19892174; http://dx.doi.org/10.1016/S0076-6879(09)63014-7.
9. Bieniossek C, Imasaki T, Takagi Y, Berger I. MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 2012; 37:49-57; PMID:22154230; http://dx.doi.org/10.1016/j.tibs.2011.10.005.
10. Bieniossek C, Nie Y, Frey D, Olieric N, Schaffitzel C, Collinson I, et al. Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods 2009; 6:447-50; PMID:19412171; http://dx.doi.org/10.1038/nmeth.1326.
11. Bieniossek C, Richmond TJ, Berger I. MultiBac: multigene baculovirus-based eukaryotic protein complex production. Curr. Protoc. Protein Sci. 2008; Chapter 5:Unit 5.20.
12. Fitzgerald DJ, Berger P, Schaffitzel C, Yamada K, Richmond TJ, Berger I. Protein complex expression by using multigene baculoviral vectors. Nat Methods 2006; 3:1021-32; PMID:17117155; http://dx.doi.org/10.1038/nmeth983.

13. Berger I, Fitzgerald DJ, Richmond TJ. Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 2004; 22:1583-7; PMID:15568020; http://dx.doi.org/10.1038/nbt1036.
14. Trowitzsch S, Bieniossek C, Nie Y, Garzoni F, Berger I. New baculovirus expression tools for recombinant protein complex production. J Struct Biol 2010; 172:45-54; PMID:20178849; http://dx.doi.org/10.1016/j.jsb.2010.02.010.
15. Vijayachandran LS, Viola C, Garzoni F, Trowitzsch S, Bieniossek C, Chaillet M, et al. Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 2011; 175:198-208; PMID:21419851; http://dx.doi.org/10.1016/j.jsb.2011.03.007.
16. Fitzgerald DJ, Schaffitzel C, Berger P, Wellinger R, Bieniossek C, Richmond TJ, et al. Multiprotein expression strategy for structural biology of eukaryotic complexes. Structure 2007; 15:275-9; PMID:17355863; http://dx.doi.org/10.1016/j.str.2007.01.016.
17. Palmberger D, Wilson IB, Berger I, Grabherr R, Rendic D. SweetBac: a new approach for the production of mammalianised glycoproteins in insect cells. PLoS One 2012; 7:e34226; PMID:22485160; http://dx.doi.org/10.1371/journal.pone.0034226.
18. Palmberger D, Klausberger M, Berger I, Grabherr R. MultiBac turns sweet. [epub before print]. Bioengineered 2012; 4:4; PMID:23018636.
19. Hitchman RB, Possee RD, King LA. High-throughput baculovirus expression in insect cells. Methods Mol Biol 2012; 824:609-27; PMID:22160923; http://dx.doi.org/10.1007/978-1-61779-433-9_33.
20. Maeda S, Nakagawa S, Suga M, Yamashita E, Oshima A, Fujiyoshi Y, et al. Structure of the connexin 26 gap junction channel at 3.5 A resolution. Nature 2009; 458:597-602; PMID:19340074; http://dx.doi.org/10.1038/nature07869.
21. Butkevich E, Hülsmann S, Wenzel D, Shirao T, Duden R, Majoul I. Drebrin is a novel connexin-43 binding partner that links gap junctions to the submembrane cytoskeleton. Curr Biol 2004; 14:650-8; PMID:15084279; http://dx.doi.org/10.1016/j.cub.2004.03.063.
22. Harrison RL, Bonning BC. Comparative analysis of the genomes of Rachiplusia ou and Autographa californica multiple nucleopolyhedroviruses. J Gen Virol 2003; 84:1827-42; PMID:12810877; http://dx.doi.org/10.1099/vir.0.19146-0.
23. Pijlman GP, van den Born E, Martens DE, Vlak JM. Autographa californica baculoviruses with large genomic deletions are rapidly generated in infected insect cells. Virology 2001; 283:132-8; PMID:11312669; http://dx.doi.org/10.1006/viro.2001.0854.
24. Pijlman GP, Dortmans JC, Vermeesch AM, Yang K, Martens DE, Goldbach RW, et al. Pivotal role of the non-hr origin of DNA replication in the genesis of defective interfering baculoviruses. J Virol 2002; 76:5605-11; PMID:11991989; http://dx.doi.org/10.1128/JVI.76.11.5605-5611.2002.
25. Pijlman GP, Pruijssers AJ, Vlak JM. Identification of pif-2, a third conserved baculovirus gene required for per os infection of insects. J Gen Virol 2003; 84:2041-9; PMID:12867634; http://dx.doi.org/10.1099/vir.0.19133-0.
26. Pijlman GP, Roode EC, Fan X, Roberts LO, Belsham GJ, Vlak JM, et al. Stabilized baculovirus vector expressing a heterologous gene and GP64 from a single bicistronic transcript. J Biotechnol 2006; 123:13-21; PMID:16364483; http://dx.doi.org/10.1016/j.jbiotec.2005.10.022.
27. Pijlman GP, van Schijndel JE, Vlak JM. Spontaneous excision of BAC vector sequences from bacmid-derived baculovirus expression vectors upon passage in insect cells. J Gen Virol 2003; 84:2669-78; PMID:13679600; http://dx.doi.org/10.1099/vir.0.19438-0.
28. Pijlman GP, Vermeesch AM, Vlak JM. Cell line-specific accumulation of the baculovirus non-hr origin of DNA replication in infected insect cells. J Invertebr Pathol 2003; 84:214-9; PMID:14726243; http://dx.doi.org/10.1016/j.jip.2003.10.005.
29. Carstens EB, Ye LB, Faulkner P. A point mutation in the polyhedrin gene of a baculovirus, Autographa californica MNPV, prevents crystallization of occlusion bodies. J Gen Virol 1987; 68:901-5; PMID:19731454; http://dx.doi.org/10.1099/0022-1317-68-3-901.
30. Roelvink PW, van Meer MM, de Kort CA, Possee RD, Hammock BD, Vlak JM. Dissimilar expression of Autographa californica multiple nucleocapsid nuclear polyhedrosis virus polyhedrin and p10 genes. J Gen Virol 1992; 73:1481-9; PMID:1607866; http://dx.doi.org/10.1099/0022-1317-73-6-1481.
31. Pennock GD, Shoemaker C, Miller LK. Strong and regulated expression of Escherichia coli beta-galactosidase in insect cells with a baculovirus vector. Mol Cell Biol 1984; 4:399-406; PMID:6325875.
32. Smith GE, Summers MD, Fraser MJ. Production of human beta interferon in insect cells infected with a baculovirus expression vector. Mol Cell Biol 1983; 3:2156-65; PMID:6318086.
33. Benner SA, Sismour AM. Synthetic biology. Nat Rev Genet 2005; 6:533-43; PMID:15995697; http://dx.doi.org/10.1038/nrg1637.
34. Lu TK, Collins JJ. Dispersing biofilms with engineered enzymatic bacteriophage. Proc Natl Acad Sci U S A 2007; 104:11197-202; PMID:17592147; http://dx.doi.org/10.1073/pnas.0704624104.
35. Anderson JC, Clarke EJ, Arkin AP, Voigt CA. Environmentally controlled invasion of cancer cells by engineered bacteria. J Mol Biol 2006; 355:619-27; PMID:16330045; http://dx.doi.org/10.1016/j.jmb.2005.10.076.
36. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 2010; 329:52-6; PMID:20488990; http://dx.doi.org/10.1126/science.1190719.
37. Hutchison CA 3rd, Smith HO, Pfannkoch C, Venter JC. Cell-free cloning using phi29 DNA polymerase. Proc Natl Acad Sci U S A 2005; 102:17332-6; PMID:16286637; http://dx.doi.org/10.1073/pnas.0508809102.
38. Zhao Y, Chapman DA, Jones IM. Improving baculovirus recombination. Nucleic Acids Res 2003; 31:E6-6; PMID:12527795; http://dx.doi.org/10.1093/nar/gng006.
39. Giri L, Feiss MG, Bonning BC, Murhammer DW. Production of baculovirus defective interfering particles during serial passage is delayed by removing transposon target sites in fp25k. J Gen Virol 2012; 93:389-99; PMID:21994323; http://dx.doi.org/10.1099/vir.0.036566-0.
40. Jarvis DL. Developing baculovirus-insect cell expression systems for humanized recombinant glycoprotein production. Virology 2003; 310:1-7; PMID:12788624; http://dx.doi.org/10.1016/S0042-6822(03)00120-X.
41. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10; PMID:2231712

PUBLICATION 5

The MultiBac Baculovirus/Insect Cell Expression Vector System for Producing Complex Protein Biologics.

Authors: *Duygu Sari,* ***Kapil Gupta****, Deepak Balaji Thimiri Govinda Raj, Alice Aubert, Petra Drncová, Frederic Garzoni, Daniel Fitzgerald and Imre Berger*

Adv. Exp. Med. Biol., Springer Press, New York (in press, 2015).

This manuscript presents methods involved in multiprotein complex production in insect cells using baculovirus based MultiBac system.

The MultiBac Baculovirus / Insect Cell Expression Vector System for Producing Complex Protein Biologics

AUTHORS and AFFILIATIONS

Duygu Sari[1,2], Kapil Gupta[1,2], Deepak Balaji Thimiri Govinda Raj[1,2], Alice Aubert[1,2], Petra Drncová[1,2], Frederic Garzoni[1,2], Daniel Fitzgerald[3] and Imre Berger[1,2,4]

[1] European Molecular Biology Laboratory, Grenoble Outstation, 71 Avenue des Martyrs, 38042 Grenoble Cedex 9, France

[2] Unit of Virus Host-Cell Interactions, Univ. Grenoble Alpes-EMBL-CNRS, UMI 3265, 71 Avenue des Martyrs, 38042 Grenoble Cedex 9, France

[3] Geneva Biotech SARL, Avenue de la Roseraie 64, 1205 Genève, Switzerland

[4] School of Biochemistry, University of Bristol, Bristol BS8 1TD, United Kingdom

Corresponding Author: Imre Berger (iberger@embl.fr)

Keywords: Protein complexes, Baculovirus / insect cell expression system**,** MultiBac, Structural biology, X-ray crystallography, Electron microscopy, Virus-like particles (VLPs), Vaccines, Gene transfer, Lepidoptera, Synthetic biology

ABSTRACT

Multiprotein complexes regulate most if not all cellular functions. Elucidating the structure and function of these complex cellular machines is essential for understanding biology. Moreover, multiprotein complexes by themselves constitute powerful reagents as biologics for the prevention and treatment of human diseases. Recombinant production by the baculovirus / insect cell expression system is particularly useful for expressing proteins of eukaryotic origin and their complexes. MultiBac, an advanced baculovirus / insect cell system, has been widely adopted in the last decade to produce multiprotein complexes with many subunits that were hitherto inaccessible, for academic and industrial research and development. The MultiBac system, its development and numerous applications are presented. Future opportunities for utilizing MultiBac to catalyze discovery are outlined.

1/ Introduction –The baculovirus expression vector system (BEVS)

More than 30 years ago, the high level production of a heterologous protein by using an insect specific baculovirus, derived from the *Autographa californica multiple nuclear polyhedrosis virus* (AcMNPV) was reported. Max Summers and co-workers produced functional human IFN-β in insect cells infected by a recombinant baculovirus [1]. This development was made possible by the previous observations that late in its viral life cycle, baculovirus expresses at very high levels a protein, polyhedrin, which is not essential in laboratory culture. Substitution of the *polyhedrin* gene in the baculoviral polh locus by a foreign gene of interest resulted in comparably high-level expression of the desired gene product, driven by the polh promoter, without compromising virus infectivity and the viral life cycle [2]. Shortly after, a second study by Lois Miller and colleagues demonstrated that another very late promoter p10, showed similar characteristics and could also be used for high-level production of heterologous proteins [3].

These two seminal studies established the baculovirus / insect cell expression system as a powerful means to produce proteins recombinantly. In the three decades since these hallmark contributions, baculovirus expression has become a widely adopted technology for academic and industrial applications, in research and development as well as manufacturing, and a wide range of proteins have been made by baculovirus expression vector systems (BEVS) [2, 4-7]. Multicomponent virus like particles (VLPs) resembling complex virus shells, have been produced successfully with BEVS, including VLPs from bluetongue, rotavirus and others [8-11]. More recently, the first baculovirus produced proteins have been approved in the therapy or prevention of human disease, including vaccines against influenza (Flublok®) and cervical cancer (Cervarix®), and immune-therapeutics against tumors of the prostate (Provenge®) [4]. Moreover, baculovirus itself has emerged as a versatile tool for

gene therapy, either as a production system for recombinant adeno-associated viruses [12-14] or as a DNA-based gene delivery vehicle in its own right [15, 16].

The development of BEVS for heterologous protein production and its manifold exploits have been authoritatively reviewed recently in a number of contributions, comprehensively recapitulating the technical aspects of this technology [2, 4, 5]. The subject of this present contribution is MultiBac, a particular baculovirus expression vector system developed and implemented more recently [17-25]. MultiBac was originally conceived to meet the imposing challenge of producing eukaryotic multiprotein complexes, vital cornerstones of biological activity, in high quality and quantity for high-resolution structural and functional analysis. The system has been uniquely successful in catalyzing multiprotein complex research globally. MultiBac, its ongoing development, its numerous applications and future prospects are reviewed in the following.

2/ The MultiBac system for expressing eukaryotic multiprotein complexes

Protein complexes catalyze key functions in the cell, and as a consequence, are an intense focus of contemporary biological research efforts. Genomics and proteomics studies have underpinned that most if not all proteins in eukaryotic cells are parts of larger assemblies, which in humans often comprise ten or more individual subunits. The complex interplay of proteins in these complexes is essential for cell homeostasis, biological activity and development. High-resolution functional and structural characterization of the large number of multiprotein assemblies in the cell is critical to understanding cell biology [20, 26, 27].

Multisubunit complexes may be purified from their native cell environment and their structure and function analyzed successfully at near-atomic resolution, provided they are sufficiently abundant and homogeneous. Well-known examples include RNA polymerases and ribosomes [28-31]. The overwhelming majority of multiprotein complexes in the cell, however, are characterized by low or very low abundance, which considerably complicates or even rules out their purification from native source material. Furthermore, it is becoming increasingly clear that many proteins may exist not only in one, but a number of distinct complexes, carrying out diverse functions depending on the partner molecules they associate with at a given time. Together, this often obstructs obtaining compositionally pure and homogeneous material by classical fractionation of cells and subsequent biochemical purification, notwithstanding significant progress notably in endogenous tagging methods to genomically modify endogenous proteins by powerful extraction aids such as tandem affinity purification (TAP) tags, for instance [32-34]. A solution to these issues is recombinant overproduction, enabled by the development and implementation of powerful overexpression technologies that can achieve high level production of homogeneous and active eukaryotic complexes for detailed mechanistic analysis at the molecular level.

Recombinant protein overproduction had a profound and game-changing impact on protein science, making previously inaccessible targets readily available. A very large number of proteins, their mutants and variants have been produced recombinantly, and their structure and function determined at high resolution. The availability of entire genomes has made it possible to address the protein repertoire of cells on a system-wide scale, applying high-throughput technologies [35]. Recombinant protein expression in *E.coli* as a prokaryotic expression host has become prevalent in molecular biology laboratories world-wide. The recombinant production of protein complexes of eukaryotic origin, however, poses a number of challenges which frequently rule out prokaryotic expression hosts. Eukaryotic proteins are often large and can exceed the size range *E.coli* can overproduce efficiently (typically up to ~100kDa). Post-translational modifications and processing are commonplace in eukaryotic proteins and can be essential for activity, but are generally not supported by a prokaryotic host. The eukaryotic protein folding machinery differs significantly from the chaperone system in *E.coli*, further restricting its utility for eukaryotic protein production. Much effort has been and is being devoted to improving prokaryotic host systems for heterologous production [36-38]. However, in many cases eukaryotic proteins and their complexes will likely require a eukaryotic expression host system for their overproduction, and if a eukaryotic system can be applied with comparable ease as *E.coli* based expression, then this system will likely be a preferred choice. The MultiBac system has been developed precisely with this intention to put in place such a eukaryotic expression system that supports high-level and high-quality production of eukaryotic proteins and their complexes, by using standard operating protocols (SOPs) which make its application comparably facile and routine as *E.coli*-based expression [18, 39-41].

2.1/ MultiBac developments

The baculovirus / insect cell expression system is particularly well-suited for the production of eukaryotic proteins. At the core of this expression technology is a recombinant baculovirus into which the heterologous genes of interest have been inserted. This composite baculovirus is then used to infect insect cell cultures grown in the laboratory. MultiBac is a more recent baculovirus / insect cell system which has been specifically tailored for the overproduction of eukaryotic complexes that contain many subunits [40]. An important prerequisite for the efficient expression of eukaryotic proteins and their complexes are easy-to-use reagents for (multi)gene assembly and delivery. Equally required are robust and standardized protocols for all steps involved in the expression experiment, from gene to purified protein complex. These steps should ideally be implemented as standard operating procedures (SOPs), especially in laboratories where the expression experiment itself and its optimizations are not the primary objective, but rather the protein complex and the determination of its structure and mechanism within a reasonable time-frame. The implementation of such SOPs will then enable non-specialist users to apply the technology with relative ease. The MultiBac system has been designed to meet these requirements [18, 40, 41].

MultiBac consists of an engineered baculovirus that has been optimized for multiprotein complex expression [17] (Fig. 1). The MultiBac baculovirus exists as a bacterial artificial chromosome (BAC) in *E.coli* cells (DH10MultiBac or DH10MB). The replicon (F-factor) present on the BAC restricts its copy number to (typically) one [42]. The MultiBac genome has been modified by deleting proteolytic and apoptotic functionalities from the baculoviral genome that were found to be detrimental for the quality of the heterologous

target complexes produced [17-19, 23, 41]. The MultiBac system furthermore comprises an array of small custom-designed DNA plasmid modules that facilitate the assembly of multigene expression cassettes and their integration into the baculoviral genome (Fig. 1). Integration of the multigene expression cassette constructions into the baculoviral genome occurs via two sites (Fig. 1). One is a mini-Tn7 attachment site embedded in a LacZα gene that is used for blue/white selection and is accessed by the Tn7 transposase which is expressed in the DH10MB cells from a helper plasmid as described previously [43]. Upon integration into this Tn7 site, the LacZα gene is disrupted; white colonies indicate successful transposition. A second entry site is formed by a short imperfect inverted repeat, LoxP, at a location distal from the Tn7 attachment site (Fig. 1). It can be accessed by means of the Cre enzyme, a site-specific recombinase that targets the LoxP imperfect inverted repeat (Fig. 1). Cre integration occurs by fusing LoxP sites present on the Multibac genome on the one hand, and on a DNA plasmid module on the other. Successful Tn7 and also Cre integration is imposed by antibiotic selection against the resistance markers encoded by the DNA plasmid modules integrated into the MultiBac genome. The integration sites can be used to integrate genes encoding for one or several multiprotein complexes of choice, but also for additional functionalities that may be required to activate or inactivate the complex (kinases, phosphatases, acetylases, deacetylases, others), support its folding (chaperones) or post-translational processing such as glycosylation [19, 23, 44-46].

The composite MultiBac baculoviral genome which contains all desired heterologous genes is then purified from small bacterial cultures using standard alkaline lysis protocols and applied to small insect cell cultures, typically in six-well plates, with a lipidic or non-lipidic transfection reagent [24, 41]. The resulting live baculovirions are harvested and applied to larger insect cell cultures for heterologous protein production and purification.

Production baculovirus is then stored for example at 4°C in the dark to avoid degeneration of viral titers. A more secure long term storage method is provided by freezing small aliquots of baculovirus infected insect cells (BIICs) that are then stored in liquid nitrogen [47].

A centerpiece of the MultiBac system is the facilitated assembly of genes into multiexpression cassettes (Fig. 1, 2). Originally, this was addressed by creating two different DNA plasmid modules that contained a so-called Multiplication Module. This module allowed iterative assembly of single or dual expression cassettes, each fitted with a promoter (p10 or polh) by restriction/ligation utilizing compatible sites that would be destroyed upon ligation [17]. This functional unit-based plasmid configuration later was popularized as 'BioBrick' in the context of synthetic biology.

One plasmid (pFBDM) accessed the Tn7 site, the second plasmid (pUCDM) accessed the LoxP site on the MultiBac genome. Both plasmids could be fitted with one to many foreign genes by means of multiplication. The pUCDM plasmid contains a conditional origin of replication derived from the phage R6Kγ. Its survival therefore hinges on the presence of a particular gene (*pir*) in the genome of specific *E.coli* strains. If a pUCDM plasmid is transformed into DH10MB cells which are *pir* negative, it will only survive if it is productively fused to the MultiBac genome by Cre-LoxP reaction [17].

The plasmid module repertoire of the MultiBac system was subsequently expanded by fitting out the pFBDM plasmid with a LoxP site, giving rise to pFL. A further plasmid was created, pKL, which in contrast to the high-copy number pFL plasmid is propagated at low copy numbers, and thus could accommodate difficult or very large genes that had turned out to be unstable in pFL. A new version of pUCDM was designed with a different resistance marker (spectinomycin), pSPL. pSPL or pUCDM could now be fused either with the MultiBac genome *in vivo* in DH10MB if Cre was expressed, or *in vitro* with pFL or pKL. All

plasmids retained the Multiplication Module and therefore could be outfitted with several to many genes in an iterative way. pUCDM and pSPL were now denoted Donor plasmids (D), while pFL and pKL were denoted Acceptors (A) (Fig. 2) [18, 41]. One Acceptor could be fused with one or two Donors by Cre-LoxP reaction *in vitro*, and productive fusions were identified by the combination of resistance markers present on the fusion. The fused AD or ADD constructs carrying several to many heterologous genes, could then be inserted into the Tn7 attachment site of the MultiBac genome by transposition in DH10MB cells as before. *In vitro* fusion of AD or ADD plasmids prior to integration into the Multibac genome via the Tn7 site did not rule out use of the LoxP site present on the viral backbone to additionally incorporate a Donor. Integration into the viral LoxP site simply had to precede the transposition reaction [18].

A *sine qua non* of contemporary structural biology is automation, to increase the through-put of expression experiments by robotic approaches. As a consequence, the multigene assembly technology of the MultiBac system was adapted to automation by using a liquid handling robot [24, 25]. For this, the tandem recombineering technology (TR) originally developed for prokaryotic complex expression [38] was adapted to the MultiBac system [48]. TR combines sequence-and-ligation independent gene insertion (SLIC) with Cre-LoxP fusion to generate multigene expression constructs in high-throughput in a robotic setup. Adaptation of the MultiBac plasmids to TR required subtle adjustments of the plasmids resulting in Acceptor plasmids pACEBac1 and pACEBac2 as well as Donor plasmids pIDK, pIDS and pIDC that are robotics-compatible [22, 24, 40]. Concomitantly, with the objective to simplify the monitoring of protein production by measuring fluorescence levels, a new baculovirus genome, EMBacY, was created expressing a yellow

fluorescent protein (YFP) from its backbone, with fluorescence intensity increasing in parallel with the quantity of the heterologous protein complex expressed at the same time.

All Acceptor and Donor plasmids developed for the MultiBac system over time are compatible with each other, and also with the MultiBac and EMBacY genomes (and other MultiBac genome derivatives) that were and are being developed (see also **2.3** and **2.4**).

2.2/ A MultiBac user access platform

The MultiBac system rapidly developed into a sought after tool following its introduction and original publication, compellingly underscoring the current need for an accessible expression technology for eukaryotic multiprotein complexes. The number of laboratories using MultiBac approaches a thousand at the time this review is written. Research groups in academia as well as the biotech and pharma industries are implementing MultiBac to produce their specimens of interest. Moreover, biotech spin-offs were founded based on MultiBac developments, including a preclinical vaccine development company, Redvax GmbH, which in 2015 was acquired by the global pharma enterprise, Pfizer.

The high demand for accessing this technology resulted in the establishment of a dedicated research and training platform at the European Molecular Biology Laboratory (EMBL) in the Eukaryotic Expression Facility (EEF) established at the EMBL Outstation in Grenoble (Fig. 3). The EEF has been supported since 2008 by the European Commission (EC) through infrastructure grants (P-CUBE, BioStruct-X, INSTRUCT) and by the national (French) research agency (ANR) through the Investissement d'Avenir program. More than a hundred projects per year, by local national and transnational users, including academic research projects as well as industrial contracts have been processed in the facility, covering a wide range of applications in basic and applied research and development. Academic

project access is based on the sole criteria of excellence, determined by an independent panel reviewing research proposals. The facility has implemented a SOP-based procedure for routinely and rapidly moving projects through the MultiBac platform pipeline [23, 24].

The entire process from preparing the multigene expression construct to (small-scale) purification of the specimens of interest requires around two weeks according to this protocol. A large number of constructs can be processed in parallel. Virus amplification as well as heterologous protein production is monitored by measuring the signal of yellow fluorescent protein (YFP) in small (1 million cells) aliquots withdrawn from the cell cultures at defined intervals. The YFP signal reaching a plateau indicates maximal production of the desired complex specimen and the expression culture is harvested at this point for further processing and purification. Particular care is taken to assure highest quality virus production during virus amplification – early (budded) virus is harvested throughout the amplification process to avoid accumulation of defective viral particles that would compromise virus performance. Initial virus is stored at 4°C, while production virus is stored as BIICs in liquid nitrogen [21-24] (Fig. 3).

2.3/ OmniBac: Universal multigene transfer vectors

Two approaches for foreign gene integration into the baculoviral genome dominate the field. One approach depends on the baculoviral genome present as a BAC in *E.coli* cells, and relies on gene integration by transposition catalyzed by the Tn7 transposase which is constitutively expressed from a dedicated (helper) plasmid in the cells harboring the BAC. The foreign genes are provided by transforming a transfer plasmid into these *E.coli* cells. Selection for recombinant BACs occurs based on the resistance marker that is present on the transfer plasmid and also integrated, as well as blue / white selection; productive transposase

mediated integration destroys a LacZα gene. The MultiBac system retains this strategy, extended by the option to provide in addition to the Tn7 transposase also the site-specific Cre recombinase transiently from a second dedicated helper plasmid, to fuse an additional Donor construct into a LoxP site present on the MultiBac baculoviral genome [17, 40, 41].

The alternative, original approach is based on homologous recombination, mediated by DNA regions in the transfer plasmid that are also present in the baculoviral genome, flanking the polh locus that had been inactivated by destruction of the *polh* gene. These regions of homology correspond to the open reading frames Orf1620 and lef2/603. By using this method, insertion occurs by co-transfecting the transfer plasmid and the purified baculoviral genome into insect cells. The baculoviral genome is typically linearized in the region between Orf1629 and lef2/603, to suppress formation of virus devoid of the heterologous DNA of interest. Fusion of the plasmid DNA with the baculoviral DNA to create a replicating genome is then achieved via the homologous recombination system of the insect cells. The genome is thus re-circularized, concomitantly inserting the gene of interest. Live virions are then produced that express the desired protein(s).

Both methods have advantages and disadvantages. The Tn7/BAC based integration approach is the method of choice in many laboratories, mainly due to its simplicity. Since the baculoviral genome exists as a BAC in *E.coli*, it can be, in theory, propagated indefinitely and used for many experiments after obtaining the initial aliquot. Moreover, it can also be manipulated by gene editing technologies in its *E.coli* host cell. In contrast, the linearized baculoviral DNA for homologous recombination has to be obtained from the supplier for every experiment *de novo* and cannot be propagated at will. Furthermore, the homologous recombination method is arguably somewhat more involved and may require specialist knowledge. On the other hand, this approach is more amenable to automation as compared to

the BAC-based method which is characterized by many steps some of which (such as blue / white screening) cannot be readily scripted into robotics routines. A further disadvantage of the BAC-based system may be found in the relative instability that was described for BAC-derived baculoviruses in insect cells, presumably originating from the presence of extended bacterial DNA elements (selection marker, F-replicon, LacZα gene) [4, 49], limiting its use in human applications mainly to preclinical studies [4]. In fact, baculoviruses used in commercial production to date are still being made almost exclusively by applying the classical homologous recombination technique [4].

Available BEVS all relied on either one method or the other, which were mutually exclusive, and the choice of the transfer plasmid decided which virus would be used for protein production. This situation is unsatisfactory given that both systems (BAC/Tn7-based and classical) provide unique opportunities. Moreover, numerous baculoviruses with customized functionalities have been created for both applications, each with its own merit. We therefore designed and created the "OmniBac" transfer plasmids which combine the DNA elements required for Tn7 transposition in the BAC-based system and also the homology regions for baculovirus generation following the classical method, and thereby are universally applicable (Fig. 4). These OmniBac transfer plasmids function as Acceptors in the MultiBac system (Fig. 4). They can be combined with a variety of Donors to yield multigene Acceptor-Donor fusions that can then be funneled into any baculovirus of choice [50].

2.4/ The ComplexLink polyprotein expression technology

A challenge frequently encountered when overexpressing protein complexes relates to unbalanced expression levels of the individual protein subunits that are to assemble into the

biological target superstructure. If a particular subunit is badly made in a co-expression experiment, it will limit the overall yield of the fully assembled protein complex dramatically. This challenge can be addressed within limits by co-infection approaches using several viruses, or the choice of promoters. In contrast to very late promoters such as polh and p10, earlier promoters are expressed at lower levels. However, it may be often impractical to resort to co-infection or to tuning protein expression levels by promoter choice, also due to the fact that the timing of the production of protein subunits will then be altered as well, with often unpredictable consequences. A solution to such problems can derive from observing the strategies certain viruses utilize to realize their proteome. Coronavirus, the agent that causes severe acute respiratory syndrome (SARS), produces its complete proteome from two long open reading frames (ORFs) that give rise to polyproteins in which the individual protein specimens are covalently linked. A highly specific protease, also encoded by the ORF, then liberates the individual proteins by cleaving them apart.

The ComplexLink technology implements this strategy for recombinant polyprotein production from polygenes [22, 40, 51] (Fig. 5). In ComplexLink, genes encoding for a protein complex of choice are conjoined to yield a single ORF. This ORF gives rise to a polyprotein in which the individual proteins are linked by short amino acid sequences representing a cleavage site for the NIa protease from tobacco etch virus (TEV) which is also encoded by the ORF and is the first protein produced. In addition, a cyan fluorescent marker protein (CFP) is present at the C-terminal end of the polyprotein construction. Upon translation, TEV protease liberates itself, and all other proteins including CFP by cleaving the TEV-specific proteolytic sites. The ComplexLink plasmids, pPBac, pKL-PBac and pOmni-PBac function as Acceptors in the MultiBac system and can be fused to Donors which may contain further genes encoding for polyproteins. The ComplexLink technology

proved to be highly successful in producing difficult-to-express protein complexes in high quality and quantity, including the physiological core complex of human General Transcription Factor TFIID [52-54]. A notable exploit is influenza polymerase, an important drug target to combat the flu, which has remained inaccessible for 40 years since its discovery. Influenza polymerase has been produced, for the first time, successfully using ComplexLink in conjunction with MultiBac, enabling elucidation of its structure and mechanism by X-ray crystallography at near-atomic resolution [55, 56] (Figs. 5, 6).

3/ Applications

The MultiBac system in its original form was introduced in 2004, and has become the method of choice for a wide range of applications. Primarily developed for accelerating structural biology of multiprotein complexes, it has since then been modified and improved to benefit also other fields, in basic and applied research. We have followed these developments with interest and have occasionally highlighted them in invited review articles and commentaries [40, 53, 57, 58]. In the following, we intend to summarize these developments without being exhaustive, focusing on recent exploits by researchers who adopted the system we had developed, to catalyze their research.

3.1/ Accelerating complex structural biology

The first users which implemented MultiBac were structural biologists interested in elucidating the architecture and mechanics of important multiprotein machines in cell biology at the molecular, near-atomic level. This is also the field where an impressive flurry of highest impact MultiBac-enabled contributions was achieved to date. We had highlighted some of these exciting structures, including important drug targets such as the LKB1-STRAD-MO25 complex [59], and the first structure of a nucleosome-bound chromatin remodeler, Isw1 [60] in a contribution just over two years ago in *Trends in Biochemical Sciences* [40]. In the short time since then, spanning a mere two years, a large number of new structural studies were carried out using material produced with MultiBac. A selection of these exploits is presented in Figure 6.

Landmark achievements are the recent crystal structures of influenza polymerase [55, 56]. This success was catalyzed by applying the ComplexLink and MultiBac technologies in combination, to produce this trimeric protein complex which had remained elusive for decades. The structures describe fluA and fluB variants of the polymerase bound to its RNA

ligand, and provide important structural insights into cap-snatching and RNA synthesis by this enzyme complex, opening up new avenues for pharmaceutical development to combat flu. Further crystallographic exploits include the structures of human cytoplasmic dynein-2 primed for its power stroke [61], the human argonaute-2 / RNA complex [62], the structure of the spliceosomal protein Prp8 bound to an RNA helicase, Brr2 [63, 64], and structures of PI4KIIIβ kinase complexes [65] among numerous others. Highlights achieved by using MultiBac produced material for electron microscopic studies include the structures of COP1-coated vesicles, revealing alternate coatomer conformations and interactions [66, 67], the architecture of the physiological core of human General Transcription Factor TFIID [52, 58], or the elucidation of the molecular mechanisms of the anaphase promoting complex APC/C at sub-nanometer resolution [68]. Recently, the entire human Mediator transcription factor holo-complex has been successfully assembled by using MultiBac, and functionally characterized [69]. MultiBac reagents have been incorporated into pipelines for producing membrane proteins and their complexes [70]. We anticipate that many more exciting structures of important protein assemblies will be determined in the future, by using the MultiBac system as a production tool.

3.2/ MultiBac in pharma and biotech

The baculovirus / insect cell system has had a major impact on the production of high-value protein targets, for pharmacological characterization, structure-based drug design, diagnostics, biosensor engineering and high-throughput proteomics [2, 4, 5, 71]. Notably human proteins, virus-like particles (VLPs) and vaccines have been successfully expressed by using BEVS [2, 4, 5]. Glycoproteins are sought-after biologics in the pharma and biotech sector, and insect cells have proven to be well suited for the expression of biologically active

and immunogenic specimens. The MultiBac system has been engineered to enable high-quality production of glycoproteins and their complexes. The original MultiBac baculoviral genome was already lacking the *v-cath* and *chiA* genes, which are encoding for cathepsin-like protease and chitinase, respectively. Both v-cath and chiA have been shown to be detrimental to glycoprotein production [72-74]. The glycosylation pattern of secreted proteins in insect cells differs from mammalian patterns which involve more complex N-glycans [75, 76]. These differences can have adverse effects on human glycoproteins produced in insect cells. To overcome this impediment, a new MultiBac-derived baculovirus, SweetBac, was constructed, which includes glycosyltransferases in the backbone, resulting in mammalianized glycosylation patterns of SweetBac-produced glycoproteins [23, 44, 45]. More recently, improved MultiBac variants were introduced to minimize fucosylation in insect cell derived glycoproteins to reduce binding to antibodies from the sera of patients with allergies [46].

The BEVS has demonstrated its aptitude to produce complex multicomponent assemblies such as virus-like particles (VLPs) [4, 9-11, 77]. VLPs are promising candidates for vaccination. VLPs resemble natural virus shells, but are lacking genetic material and therefore are safe and not infectious. VLPs can be proteinaceous, such as for example papilloma VLPs used to prevent cervical cancer. More recently, enveloped viruses have been successfully produced using BEVS, including influenza and chikungunya vaccine candidates [4, 77-80]. The MultiBac system was already successfully used to produce a number of VLPs including an array of papilloma serotypes [40]. Complex VLPs representing highly pathological virus strains were produced safely [81]. In particular the availability of OmniBac plasmids, which are part of the MultiBac vector suite, may provide unique opportunities for VLP vaccine development, as they are equally useful for exploratory

preclinical studies in high-throughput as well as pharmacological manufacturing, by choosing the most appropriate viral backbone for large scale expression (Fig. 4).

Baculoviruses not only infect insect cells, but can also transduce mammalian cells efficiently [15, 82-84]. By choosing mammalian-active promoters instead of polh, p10 or other baculoviral promoters, proteins of interest can be produced from a baculovirus that has entered a mammalian host cell. Baculoviruses do not replicate in mammalian cells, therefore, the current consensus is that this BacMam approach can be performed safely in laboratories. A particular benefit of BacMam is that large DNA insertions including multicomponent signaling cascades or entire metabolic pathways can be transduced into mammalian cells by the baculovirus which can tolerate very large gene insertions, and can be amplified and produced in large amount in a straight-forward manner in insect cell cultures. Baculovirus is thus emerging as a highly promising gene delivery tool into mammalian cells, for a multitude of applications [14, 15]. Already, multigene MultiBac constructions were successfully used to produce recombinant adeno-associated viruses (AAVs) for gene therapy [12, 13, 40].

3.3/ Synthetic Biology: Rewiring the genome

The AcMNPV genome has a size of around 130 kilobases and contains numerous functionalities, which are essential in the natural life cycle of the virus, but dispensable in laboratory culture. Moreover, in the laboratory, it has been recognized that the genome has a tendency to undergo multiple deletions in its genome during amplifications, notably in regions containing foreign gene cargo that has been inserted for overproduction [4, 50]. This can have severely detrimental consequences for heterologous target protein production, especially for fermenter-scale manufacturing of biologics which require large foreign DNA insertions in the viral genome and several rounds of amplification, until sufficient volumes of

production virus are obtained to charge the fermenter. We have shown that some of these limitations can be overcome at least on laboratory-scale by stringently adhering to virus generation protocols that avoid or limit the occurrence of widespread deletions [18]. Moreover, it appears that there is considerable scope for improving virus performance by eliminating mutational or deletion hot-spots in the viral genome, and by removing DNA regions which are not required in cell culture. The baculoviral genome, in particular when present as a BAC, lends itself excellently to genome manipulation and editing techniques. We and others have exploited this avenue to remove unnecessary or undesired functionalities from the virus, such as the *polh*, *p10*, *v-cath* and *chiA* genes among (many) others.

The advent of efficient synthetic biology techniques now holds the promise to reverse this somewhat cumbersome top-down approach, and to rationally redesign and rewire the baculoviral genome bottom-up by applying DNA synthesis and assembly methods that have become available more recently. Interestingly, comparison of baculovirus sequences in genome databases indicate that most genes and DNA elements thought to be required for survival of the virus in cell culture are confined to roughly one half of the circular viral genome, while the other half contains DNAs that can be probably largely disposed of in the laboratory [50]. This approach may hold challenges given the complex interplay of baculoviral proteins and their relative expression levels during the different phases of the viral life cycle [4]. Notwithstanding, synthetic approaches, probably best applied in combination with sequential deletions, are exciting and potentially highly rewarding avenues for developing new and minimal baculoviral genomes that can be customized for optimal properties in the research laboratory and also in industrial manufacturing.

4/ Outlook

Since the pioneering first reports more than three decades ago, the baculovirus / insect cell expression system has developed into a mainstream production platform, accelerating a wide range of research projects in academic and industrial laboratories. In the post-genomic era, multiprotein complexes have entered center stage as essential catalysts of cellular activity, and notably the MultiBac BEVS has contributed substantially to make hitherto inaccessible protein complexes available, to unlock their structure and mechanism in molecular detail. Moreover, BEVS has emerged as a remarkably useful tool in the biotech and pharma sector, for the production of complex biologics in disease prevention and therapy. The development of this versatile expression tool is continuing unabated as it is set to benefit markedly from powerful new synthetic biology techniques that are becoming readily available. Fueled by these innovations, BEVS is excellently positioned to play a key and increasing role in the life sciences, in basic and applied research in the future. Exciting times are ahead of us.

Acknowledgement

We thank all members of the Berger laboratory for their support, and in particular Lakshmi S. Vijayachandran (Amrita Center for Nanosciences & Molecular Medicine, India), Yan Nie (MPI, Dortmund), Simon Trowitzsch (Goethe University, Frankfurt) and Christoph Bieniossek (Hofmann-La Roche, Basel) for their contribution and helpful discussions. This work was funded by the European Commission (EC) Framework Programme (FP) 7 ComplexINC project (grant number 279039), to DF and IB. The MultiBac user access platform at the EMBL Grenoble received funding from the EC through the FP6 and FP7 projects SPINE2C, P-CUBE, BioStruct-X and INSTRUCT (grant numbers 031220, 227764, 283570 and 211252), and from the French Agence Nationale de Recherhe (ANR) through the Investissement d'Avenir funding scheme. DBTGR was a fellow of the EMBL interdisciplinary post-doctoral opportunities program (EIPOD).

References

1. Smith GE, Summers MD and Fraser MJ (1983) Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Mol Cell Biol* 3:2156-2165.
2. Summers MD (2006) Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides. *Adv Virus Res* 68:3-73.
3. Pennock GD, Shoemaker C and Miller LK (1984) Strong and regulated expression of Escherichia Coli beta-galactosidase in insect cells with a baculoviral vector. *Mol Cell Biol* 4:399-406.
4. van Oers MM, Plijman GP and Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. *J Gen Virol* 96:6-23.
5. Contreras-Gomez A, Sanchez-Miron F, Garcia-Camacho F, Molina-Grima E and Chisti Y (2014) Protein production using the baculovirus-inseect cell expression system. *Biotechnol Prog* 30:1-18.
6. Jarvis DL (2009) Baculovirus-insect cell expression systems. *Methods Enzymol* 463:191-222.
7. Possee RD and King LA (2007) Baculovirus transfer vectors. *Methods Mol Biol* 388:55-76.
8. Perez de Diego AC *et al*. (2011) Characterisation of protection afforded by a bivaleent virus-like partcile vaccine against bluetongue virus serotypes 1 and 4 in sheep. *PLoS ONE* 6:e26666.
9. Vicente T, Roldao A, Peixeto C, Carrondo MJT and Alves PM (2011) Large-sclae production and purification of VLP-based vaccines. *J Invertebr Pathol* 107(Suppl): S42-S48.
10. Roy P and Noad R (2009) Bluetongue vaccines. *Vaccine* 27(Suppl):D86-89
11. Roy P and Noad R (2009) Virus-like partciles as a vaccine delivery system: myths and facts. *Adv Exp Med Biol* 655:145-158.
12. Mietzsch M *et al*. (2014) OneBac: platform for scalable and high-titer production of adeno-associated virus serotype 1-12 vectors for gene therapy. *Hum Gene Ther* 25(3):212-222.
13. Marsic D *et al*. (2014) Vector design Tour de Force: integrating combinatorial and rational approaches to derive novel adeno-associated virus variants. *Mol Ther* 22(11):1900-1909.
14. Kotin RM (2011) Large-scale recombinant adeno-associated virus production. *Hum Mol Genet* 20(R1):R2-6.
15. Airenne KJ *et al*. (2013) Baculovirus: an insect-derived vector for diverse gene transfer applications. *Mol Ther* 21(4):739-749.

16. Paul A, Hasan A, Rodes L, Sangaralingam M and Prakash S (2014) Bioengineered baculoviruses as new class of therapeutics using micro and nanotechnologies: principles, prospects and challenges. *Adv Drug Deliv Rev* 71:115-130.

17. Berger I, Fitzgerald DJ and Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. *Nat Biotechnol* 22(12):1583-1587.

18. Fitzgerald DJ *et al*. (2006) Protein complex expression by using multigene baculoviral vectors. *Nat Methods* 3(12):1021-1032.

19. Fitzgerald DJ *et al*. (2007) Multiprotein expression strategy for structural biology of eukaryotic complexes. *Structure* 15(3):275-279.

20. Bieniossek C and Berger I (2009) Towards eukaryotic structural complexomics. *J Struct Funct Genomics* 10(1):37-46.

21. Trowitzsch S, Bieniossek C, Nie Y, Garzoni F and Berger I (2010) New baculovirus expression tools for recombinant protein complex production. *J Struct Biol*. 172(1):45-54.

22. Vijayachandran LS *et al*. (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J Struct Biol* 175(2):198-208.

23. Trowitzsch S, Palmberger D, Fitzgerald D, Takagi Y and Berger I (2012) MultiBac complexomics. *Expert Rev Proteomics* 9(4):363-73.

24. Berger I *et al*. (2013) The MultiBac protein complex production platform at the EMBL. *J Vis Exp* 11(77):e50159.

25. Berger I, Chaillet M, Garzoni F, Yau-Rose S and Zoro B (2013) High-throughput screening of multiple protein complexes. *AM Lab* 25(8):32-35.

26. Nie Y *et al*. (2009) Getting a grip on complexes. *Curr Genomics* 10(8):558-572.

27. Robinson CV, Sali A and Baumeister W (2007) The molecular sociology of the cell. *Nature* 450(7172):973-982.

28. Ramakrishnan V (2014) The ribosome emerges from a black box. *Cell* 159(5):979-984.

29. Fernández-Tornero C *et al*. (2013) Crystal structure of the 14-subunit RNA polymerase I. *Nature* 502(7473):644-649.

30. Engel C, Sainsbury S, Cheung AC, Kostrewa D and Cramer P (2013) RNA polymerase I structure and transcription regulation. *Nature* 502(7473):650-655.

31. Cramer P *et al*. (2008) Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* 37:337-352.

32. Völkel P, Le Faou P and Angrand PO (2010) Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry. *Biochem Soc Trans* 38(4):883-887.

33. Li Y (2010) Commonly used tag combinations for tandem affinity purification. *Biotechnol Appl Biochem* 55(2):73-83.

34. Janin J and Séraphin B (2003) Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol* 13(3):383-388.

35. Almo SC *et al*. (2013) Protein production from the structural genomics perspective: achievements and future needs. *Curr Opin Struct Biol* 23(3):335-344.

36. Haffke M *et al*. (2015) Characterization and Production of Protein Complexes by Co-expression in Escherichia coli. *Methods Mol Biol* 1261:63-89.

37. Vincentelli R and Romier C (2013) Expression in Escherichia coli: becoming faster and more complex. *Curr Opin Struct Biol* 23(3):326-334.

38. Bieniossek C *et al*. (2009) Automated unrestricted multigene recombineering for multiprotein complex production. *Nat Methods* 6(6):447-450.

39. Abdulrahman W *et al*. (2015) The production of multiprotein complexes in insect cells using the baculovirus expression system. *Methods Mol Biol* 1261:91-114.

40. Bieniossek C, Imasaki T, Takagi Y and Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. *Trends Biochem Sci* 37(2):49-57.

41. Bieniossek C, Richmond TJ and Berger I (2008) MultiBac: multigene baculovirus-based eukaryotic protein complex production. *Curr Protoc Protein Sci* Chapter 5:Unit 5.20.

42. Tsutsui H and Matsubara K (1981) Replication control and switch-off function as observed with a mini-F factor plasmid. *J Bacteriol* 147(2):509-516.

43. Luckow VA, Lee SC, Barry GF and Olins PO (1993) Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. *J Virol* 67:4566-4579.

44. Palmberger D, Wilson IB, Berger I, Grabherr R and Rendic D (2012) SweetBac: a new approach for the production of mammalianised glycoproteins in insect cells. *PLoS One* 7(4):e34226.

45. Palmberger D, Klausberger M, Berger I and Grabherr R (2013) MultiBac turns sweet. *Bioengineered* 4(2):78-83.

46. Palmberger D *et al*. (2014) Minimizing fucosylation in insect cell-derived glycoproteins reduces binding to IgE antibodies from the sera of patients with allergy *Biotechnol J* 9(SI)1206-1214.

47. Wasilko DJ *et al*. (2009) The titerless infected-cells preservation and scale-up (TIPS) method for large-scale production of NO-sensitive human soluble guanylate cyclase (sGC) from insect cells infected with recombinant baculovirus. *Protein Expr Purif* 65(2):122-132.

48. Haffke M, Viola C, Nie Y and Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research *Methods Mol Biol* 1073:131-140.

49. Plijman GP, van Schijndel JE and Vlak JM (2003) Spontaneous excision of BAC vector sequences from bacmid-derived baculovirus expression vectors upon passage in insect cells. *J Gen Virol* 84:2669-2678.

50. Vijachandran LS et al. (2013) Gene gymnastics: Synthetic biology for baculovirus expression vector system engineering. *Bioengineered* 4(5):279-287.

51. Nie Y, Bellon-Echeverria I, Trowitzsch S, Bieniossek C and Berger I (2014) Multiprotein complex production in insect cells by using polyproteins. *Methods Mol Biol* 1091:131-141.

52. Bieniossek C *et al*. (2013) The architecture of human general transcription factor TFIID core complex. *Nature* 493(7434):699-702.

53. Barford D, Takagi Y, Schultz P and Berger I (2013) Baculovirus expression: tackling the complexity challenge. *Curr Opin Struct Biol* 23(3):357-364.

54. Trowitzsch S *et al*. (2015) Cytoplasmic TAF2-TAF8-TAF10 complex provides evidence for nuclear holo-TFIID assembly from preformed submodules. *Nat Commun* 14;6:6011.

55. Reich S et al. (2014) Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* 516(7531):361-366.

56. Pflug A, Gulligay D, Reich S and Cusack S (2014) Structure of influenza A polymerase bound to the viral RNA promoter. *Nature* 516(7531):355-360.

57. Berger I and Mary LM (2013) Protein production for structural biology: new solutions to new challenges. *Curr Opin Struct Biol* 23(3):317-318.

58. Kandiah E, Trowitzsch S, Gupta K, Haffke M and Berger I (2014) More pieces to the puzzle: recent structural insights into class II transcription initiation. *Curr Opin Struct Biol* 24:91-97.

59. Zeqiraj E, Filippi BM, Deak M, Alessi DR and van Aalten DM (2009) Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. *Science* 326(5960):1707-1711.

60. Yamada K et al. (2011) Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature* 472(7344):448-453.

61. Schmidt H, Zalyte R, Urnavicius L and Carter AP (2015) Structure of human cytoplasmic dynein-2 primed for its power stroke. *Nature* 518(7539):435-438.

62. Elkayam E *et al*. (2012) The structure of human argonaute-2 in complex with miR-20a. *Cell* 150(1):100-110.

63. Santos KF *et al*. (2012) Structural basis for functional cooperation between tandem helicase cassettes in Brr2-mediated remodeling of the spliceosome. *Proc Natl Acad Sci USA* 109(43):17418-17423.

64. Mozaffari-Jovin S *et al*. (2013) Inhibition of RNA helicase Brr2 by the C-terminal tail of the spliceosomal protein Prp8. *Science* 341(6141):80-84.

65. Burke JE *et al*. (2014) Structures of PI4KIIIβ complexes show simultaneous recruitment of Rab11 and its effectors. *Science* 344(6187):1035-1038.

66. Sahlmuller MC *et al*. (2011) Recombinant heptameric coatomer complexes: novel tools to study isoform-specific functions. *Traffic* 12(6):682-692.

67. Faini M *et al*. (2012) The structures of COPI-coated vesicles reveal alternate coatomer conformations and interactions. *Science* 336(6087):1451-1454.

68. Chang L, Zhang Z, Yang J, McLaughlin SH and Barford D (2014) Molecular architecture and mechanism of the anaphase-promoting complex. *Nature* 513(7518):388-393.

69. Cevher MA *et al*. (2014) Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. *Nat Struct Mol Biol* 21(12):1028-1034.

70. Goehring A *et al*. (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat Protoc*. 9(11):2574-2585.

71. Drugmand JC, Schneider YJ and Agathos, SN (2012) Insect cells as factories for biomanufacturing . *Biotechnol Adv* 30(5):1140-1157.

72. Hom LG, Ohkawa T, Trudeau D and Volkman LE (2002) *Autographa californica* M nucleopolyhedrovirus ProV-CATH is activated during infected cell death. *Virology* 131:561-565.

73. Hitchman RB *et al*. (2010) Improved expression of secreted and membrane-targeted proteins in insect cells. *Biotechnol Appl Biochem* 56:85-93.

74. Kaba SA, Salceda AM, Wafula PO, Vlak JM and van Oers MM (2004) Development of a chitinase and v-cathepsin negative bacmid of improved integrity of secreted recombinant proteins. *J Virol Methods* 122:113-118.

75. Harrison RL and Jarvis DL (2006) Protein N-glycosylation in the baculovirus-insect cell system and engineering of insect cells to produce “mammalianized” recombinant glycoproteins. *Adv Virus Res* 68:159-191.

76. Harrison RL and Jarvis DL (2007) Transforming lepidopteran insect cells for improved protein processing. *Methods Mol Biol* 388:341-356.

77. Fernandes F, Teixeira AP, Carinhas N, Carrondo MJ, Alves PM (2013) Insect cells as a production platform of complex virus-like particles. *Expert Rev Vaccines* 12(2):225-236.

78,. Metz SW *et al*. (2013) Effective chickungunya virus-like particle vaccine produced in insect cells. *PLoS Negl Trop Dis* 7:e2124.

79. Smith GE *et al*. (2013) Development of influenza H7N9 virus like particle (VLP) vaccine: homologous A/Anhui/1/2013 (H7N9) protection and heterologous A/chicken/Jalisco/CPA/2012 (H7N3) cross-protection in vaccinated mice challenged with H7N9 virus. *Vaccine* 31:4305-4313.

80. Metz SW and Plijman GP (2011) Arbovirus vaccines: opportunities for the baculovirus-insect cell expression system. *J Invertebr Pathol* 107(Suppl):S16-S30.

81. Behzadian F *et al*. (2013) Baculoviral Co-Expression of HA, NA and M1 Proteins of Highly Pathogenic H5N1 Influenza Virus in Insect Cells. Jundishapur J Microbiol. 6(9)e7665.

82. Condreay JP and Kost TA (2007) Baculovirus expression vectors for insect and mammalian cells. *Curr Drug Targets* 8(10):1126-1131.

83. Ames RS, Kost TA and Condreay JP (2007) BacMam technology and its application to drug discovery. *Expert Opin Drug Discov* 2(12):1669-1681.

84. Kost TA, Condreay JP and Ames RS (2010) Baculovirus gene delivery: a flexible assay development tool. *Curr Gene Ther* 10(3):168-173.

Figures Legends

Figure 1. **The MultiBac baculovirus / insect cell expression system**. The MultiBac system is shown in a schematic view (left). MultiBac consist of an engineered baculovirus optimized for protein complex production. This MultiBac baculoviral genome exists as a bacterial artificial chromosome (BAC) in *E.coli* cells. It contains two integration sites for foreign genes, by Tn7 transposition or, alternatively, by site-specific recombination mediated by the Cre enzyme. MultiBac further consists of an array of plasmids called Acceptors and Donors that facilitate multigene assembly. MultiBac baculovirions (centre) are generated by transfecting composite MultiBac BAC in insect cells. MultiBac is successfully used for a wide variety of applications in basic and applied research and development (right).

Figure 2. **MultiBac tool-kits**. A variety of entry plasmids to integrate heterologous genes into the MultiBac baculoviral genome have been created since the introduction of the system in 2004, each with its own merits. Functional modules contained in the plasmids are listed (bottom). All plasmids are compatible with each other and can be used in various combinations to generate recombinant MultiBac baculoviral genomes for multiprotein expression and/or multigene delivery. Expression cassettes are in 'BioBrick' format and enable iterative multiplication.

Figure 3. MultiBac expression platform at the EMBL Grenoble. The standard operating protocol (SOP) implemented is illustrated, to express proteins and their complexes by MultiBac. The entire process takes two weeks from generation of the composite MultiBac BAC (bacmid) to quantitative expression analysis. A MultiBac baculovirus variant called EMBacY is used in the platform, producing yellow fluorescent protein (YFP) to track virus performance and heterologous protein production. In addition, protein production is

monitored by Western blot (WB) analysis or by gel electrophoresis (SDS-PAGE). Production virus is stored long-term in frozen aliquots of baculovirus in insect cells (BIICs).

Figure 4. **OmniBac – Universal transfer plasmids for every BEVS.** (a) Acceptor plasmids pOmniBac1 and pOmniBac2 are shown schematically, functional modules are same as listed above (Fig. 2, bottom). These Acceptors can be combined with the Donor plasmids by Cre-LoxP recombination. (b) OmniBac plasmids comprise elements for homologous recombination as well as Tn7-based transposition. Multigene constructions based on OmniBac plasmids can therefore access all available baculovirus genomes. Thus, with the same plasmids, composite baculovirus for preclinical studies as well as manufacturing can be produced efficiently.

Figure 5. **ComplexLink technology**. (a) The ComplexLink technology was created to produce multiprotein complexes from self-processing polyprotein constructs as shown here schematically. The polyprotein contains a TEV protease and a fluorescent protein, at the N- and C-termini, respectively. Polyproteins are processed into the individual protein entities by the highly specific TEV protease. (b) Polyprotein expression plasmids pPBac, pKL-PBac and pOmni-PBac are shown. DNA modules are marked as above (Fig 2, bottom). (c) Schematic representation of a self-processing polyprotein encoding for influenza polymerase, before TEV-mediated cleavage. TEV stands for tobacco etch virus NIa protease, PA, PB1 and PB2 are subunits of the trimeric influenza enzyme complex, CFP stands for cyan fluorescent protein. BstEII and RsrII are asymmetric restriction enzymes that can be used to access polyprotein expression plasmids for restriction / ligation-based heterologous gene integration.

Figure 6. **MultiBac complex structure gallery.** A selection of recent high impact structures of MultiBac-produced biological specimens are shown. Examples include cryo-EM

architectures of COPI-coated Vesicles (EMD-2084 to EMD-2088), the complete human APC/C complex with coactivator and substrate (EMD-2651 to EMD-2654) and the human core-TFIID complex (EMD-2229 to EMD-2231). Notable structures that were determined by X-ray crystallography (PDB identifiers are provided in in brackets) include influenza polymerases A and B bound to the viral RNA promoter (PDB identifier 4WSA, 4WRT), human Argonaute Ago2 in complex with miR-20a RNA (4F3T), the spliceosomal complex $Brr2^{HR}$ / $Prp8^{Jab1}$ (4KIT), human GABA(B) receptor (PDB 4MQE), the dynein-2 motor bound to ADP (4RH7), the mitotic checkpoint complex MCC (4AEZ) and the GluN1/GluN2B *N*-methyl-D-aspartate receptor (3QEL). Molecular illustrations were prepared with PyMOL (www.pymol.org) and Adobe Photoshop Version CS6.

Figures

Figure 1. **The MultiBac baculovirus / insect cell expression system**.

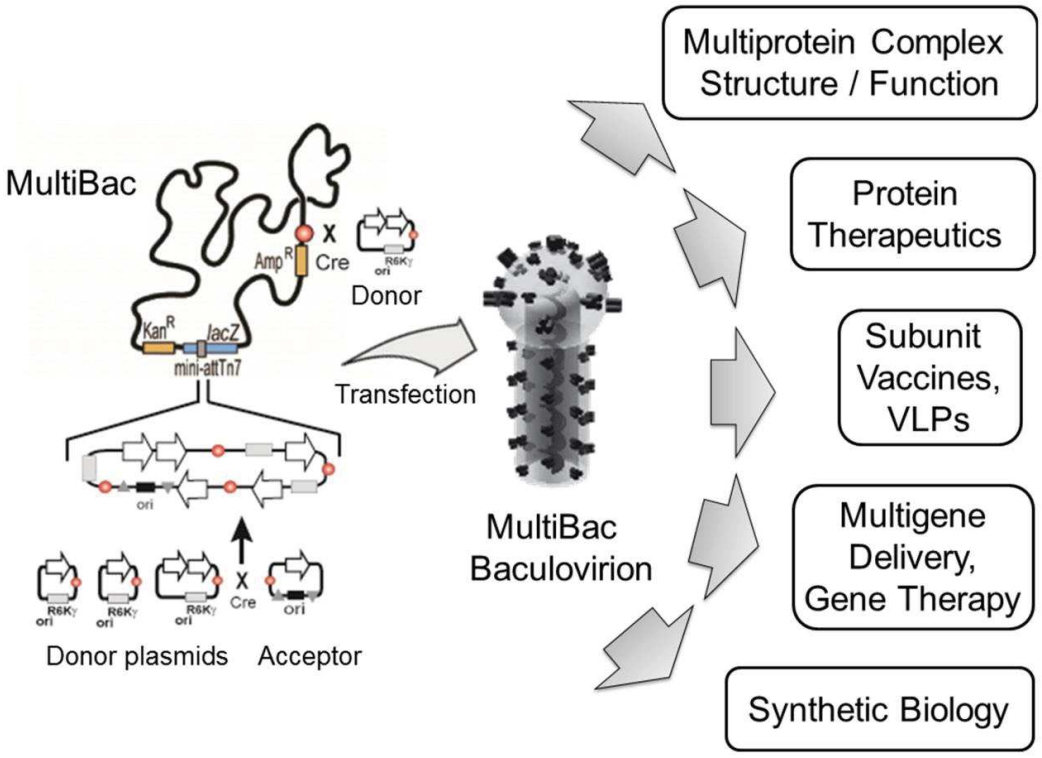


Figure 2. **MultiBac Tool-kits**.

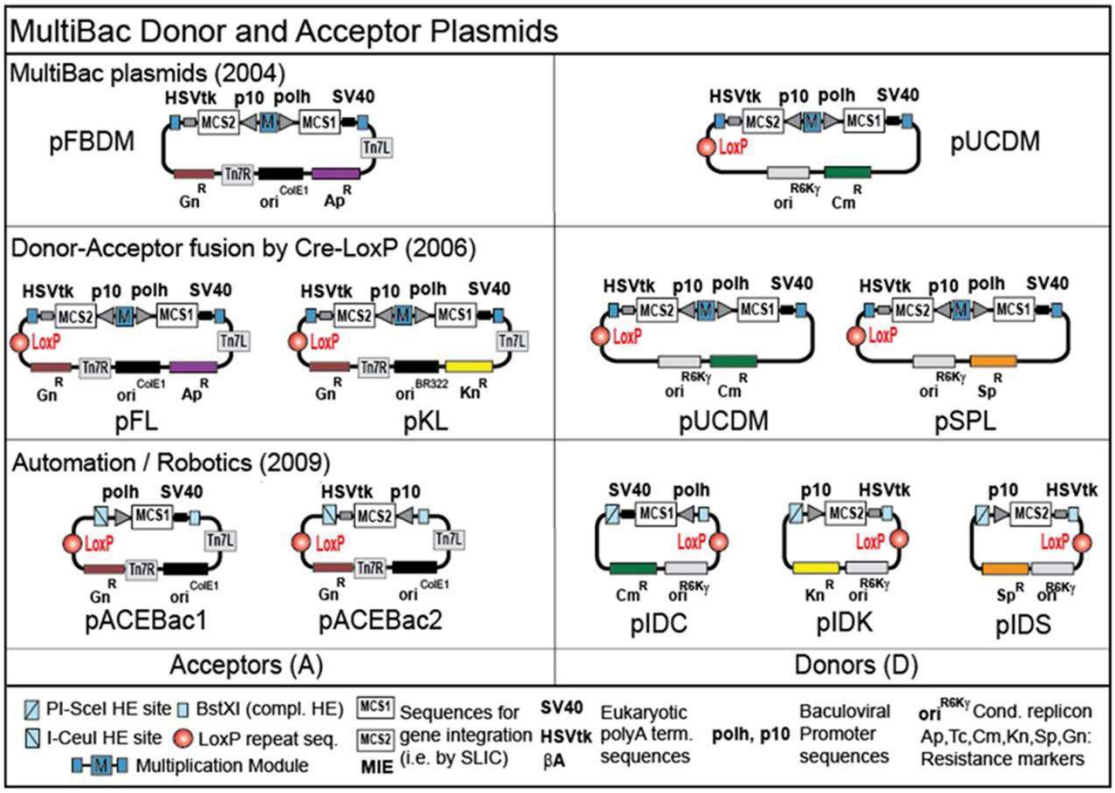


Figure 3: MultiBac expression platform at the EMBL Grenoble.

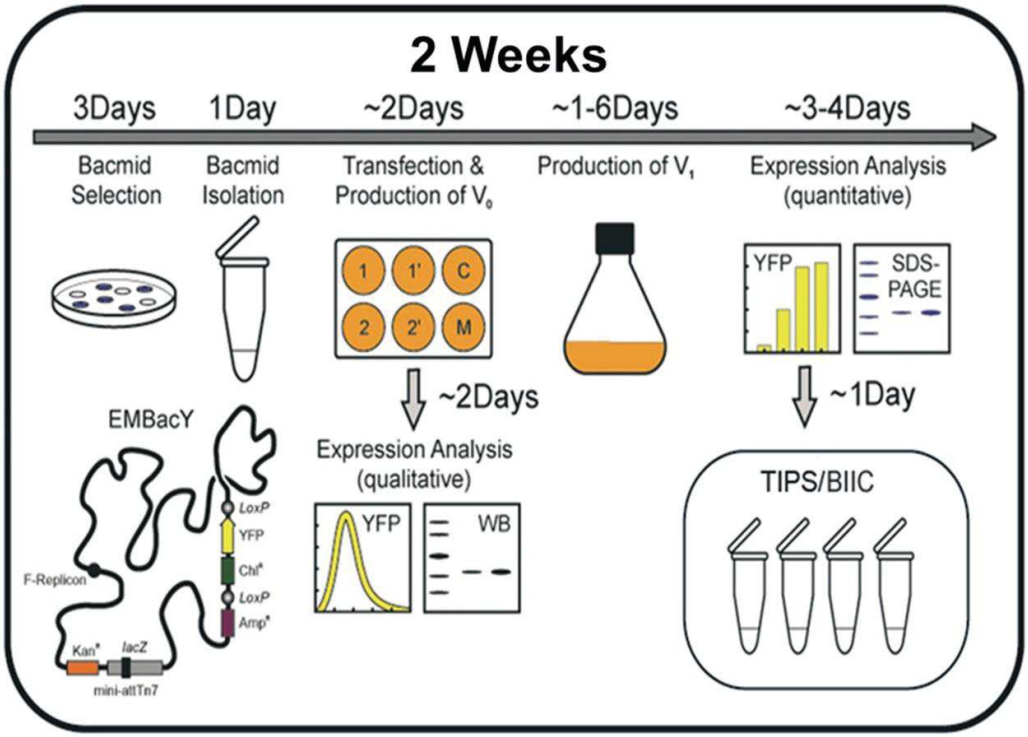


Figure 4. **OmniBac – Universal transfer plasmids for every BEVS.**

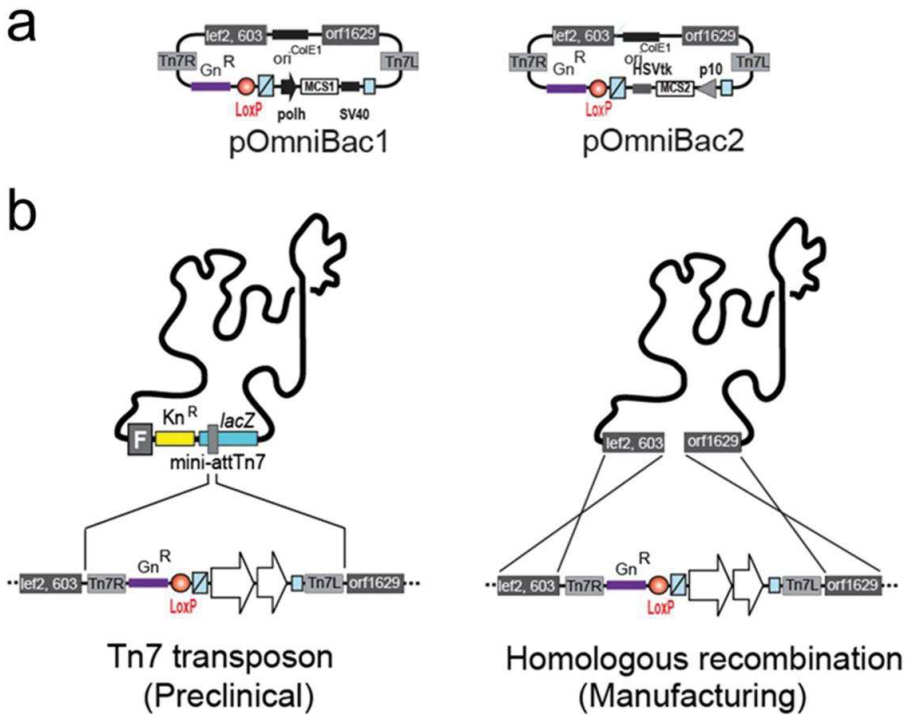


Figure 5. **ComplexLink technology.**

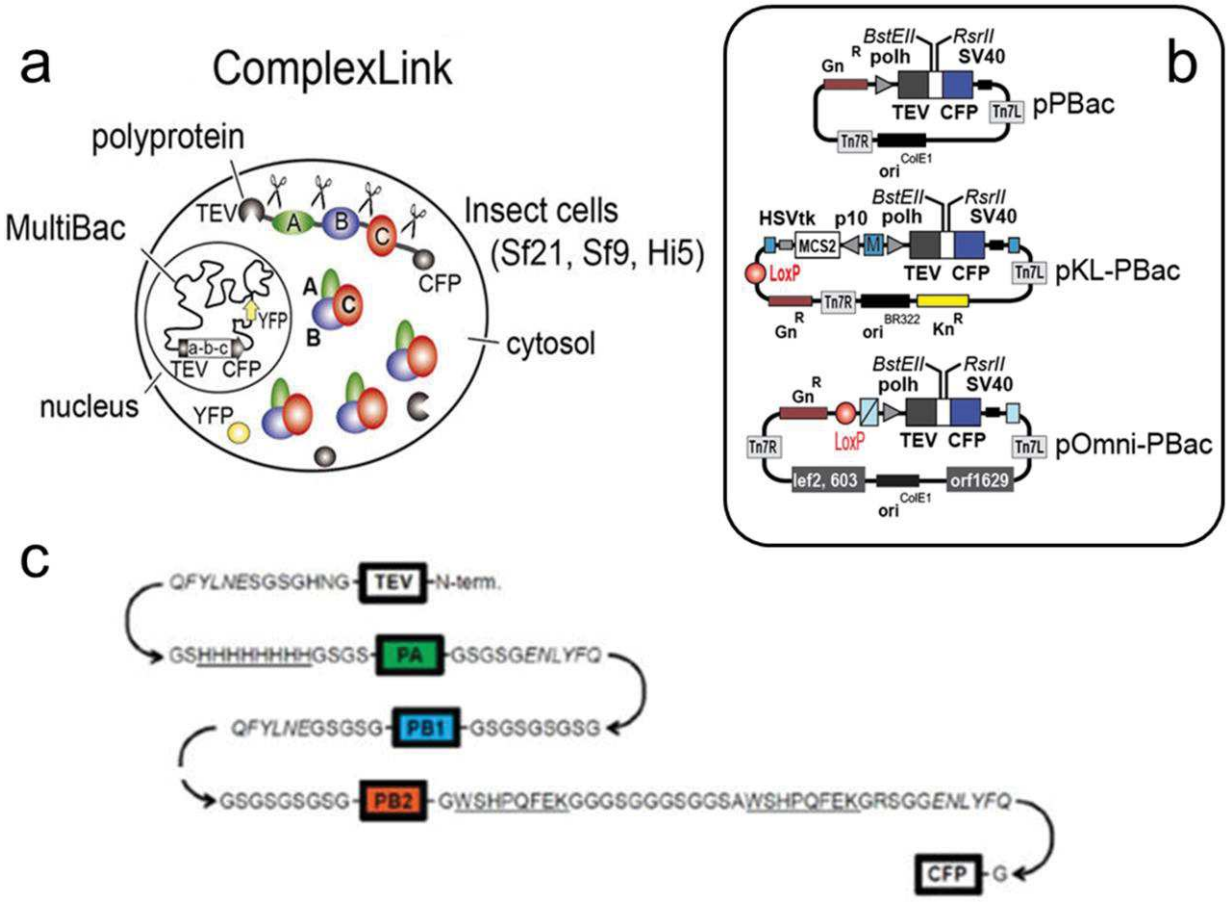
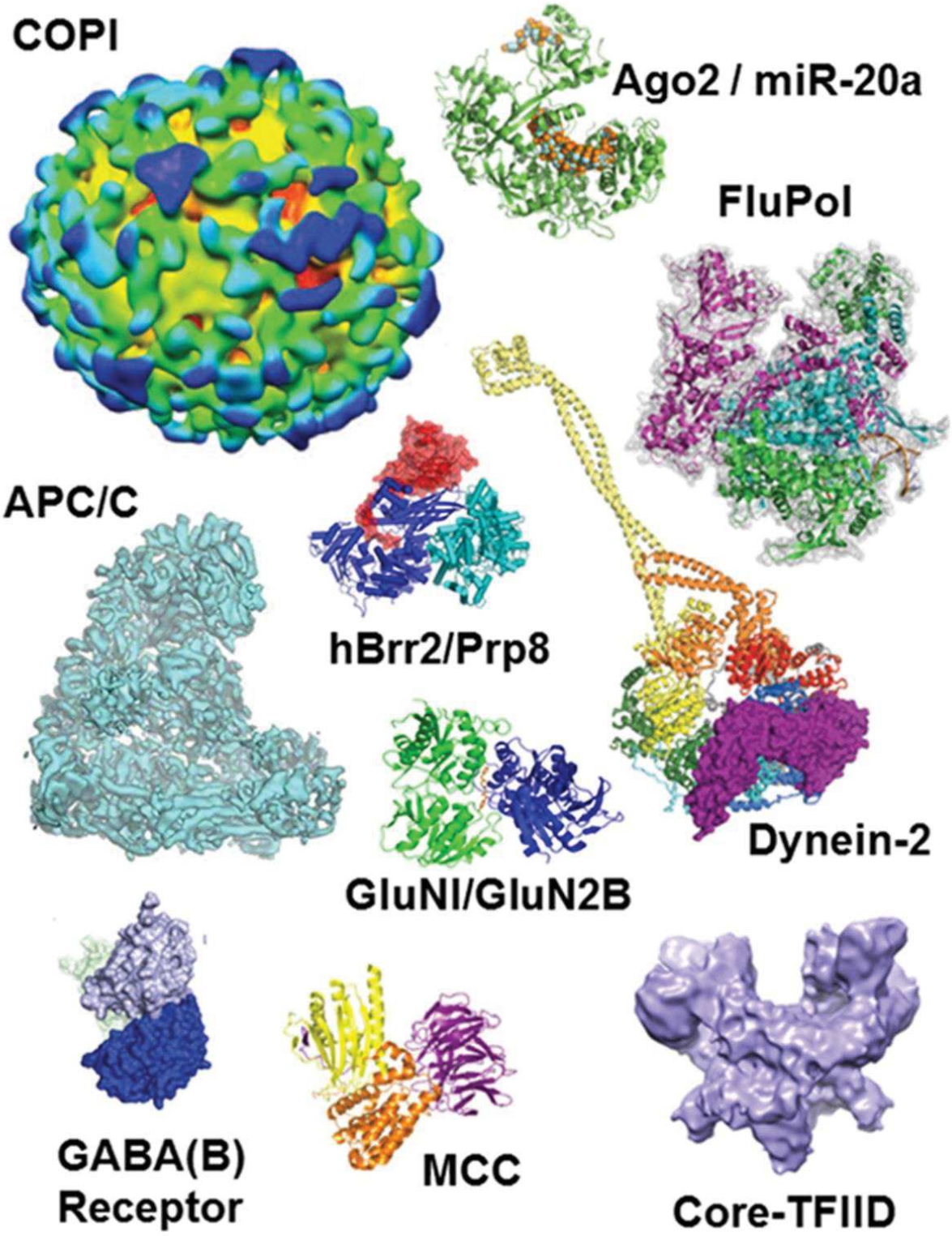


Figure 6. **MultiBac complex structure gallery**.



3. RESULTS

SUMMARY IN ENGLISH

In this chapter, we describe the results obtained to fulfil the aim of this study. Here, structural and functional characterizations of three different human TFIID subassemblies are presented. The complete human TAF1/TAF7 interaction domain complex, a novel TAF11/TAF13/TBP complex, and the 9TAF subcomplex were analyzed.

First, characterization of the structure and dynamics of the complete interaction interface of human TAF1/TAF7 complex was achieved by hybrid methods, combining the crystal structure determination of a $TAF1^{578-1106}$/$TAF7^{1-193}$ complex at 1.75 Å resolution with determining the NMR solution structure of $TAF1^{1157-1207}$/$TAF7^{153-190}$ complex, and integrating these results with data derived from SAXS analysis of the $TAF1^{578-1210}$/$TAF7^{1-193}$ and $TAF1^{578-1106}$/$TAF7^{1-193}$ complexes, to reveal the complete human TAF1/TAF7 interaction complex.

Next, analysis of a novel TAF11/TAF13/TBP complex which we discovered is described. We found this hitherto unobserved, stable trimeric complex when we were attempting to prepare the pentameric TAF11/TAF13/TFIIA/TBP/DNA complex that had been suggested by others previously based on genetic and biochemical data. This analysis includes EMSA, SEC analysis, native MS, AUC, MBP-pulldowns, SAXS analysis and CLMS analysis to reveal the characteristics of this heterotrimeric complex which may constitute a novel case of TATA-box mimicry. SAXS and CLMS analysis also provided preliminary structural information of this complex.

Finally, the negative stain EM random conical tilt (RCT) analysis of a large TFIID subassembly, called 9TAF complex, is presented.

RÉSUMÉ EN FRANÇAIS

Dans ce chapitre, nous décrivons les résultats obtenus tout au long de notre étude. Ici, la caractérisation structurelle et fonctionnelle des trois sous-ensembles de TFIID humains différents sont présentés. Le domaine d'interaction humain complet du complexe TAF1/TAF7, un nouveau complexe TAF11/TAF13/TBP et le sous-complexe 9TAF ont été analysés.

Tout d'abord, la caractérisation de la structure et la dynamique de l'interface d'interaction complète du complexe TAF1/TAF7 humain a été obtenue par des procédés hybrides, combinant la détermination de la structure cristallisée du complexe $TAF1^{578-1106}$/$TAF7^{1-193}$ à une résolution de 1,75Å avec la détermination de la structure en solution par RMN du complexe $TAF1^{1157-1207}$/$TAF7^{153-190}$, l'intégration de ces résultats avec les données issues de l'analyse SAXS du complexe $TAF1^{578-1210}$/$TAF7^{1-193}$ et du complexe $TAF1^{578-1106}$/$TAF7^{1-193}$.

Ensuite, l'analyse d'un nouveau complexe TAF11/TAF13/TBP que nous avons découvert est décrit. Nous avons découvert un complexe passé inaperçu jusqu'ici, un trimère stable lors de la préparation du pentamère TAF11/TAF13/TFIIA/TBP/ADN. Ce dernier avait été suggéré auparavant basée sur des données génétiques et biochimiques. Cette analyse comprend l'EMSA, l'analyse de la SEC, MS native, AUC, MBP-pull-downs, et l'analyse SAXS CLMS pour révéler les characterisstics de ce complexe hétérotrimérique peuvent constituer un nouveau cas de mimétisme de la boîte TATA. SAXS et CLMS ont en plus apporté des informations structurelles préliminaire sur ce complexe.

Enfin, l'analyse d'un grand sous-ensemble de TFIID, appelé complexe 9TAF par EM en coloration négative aléatoire à inclinaison conique (RCT), est présenté.

3.1. Structural characterization of a human TAF1/TAF7 complex

The interaction of the two TFIID subunits TAF1 and TAF7 was suggested based on yeast two hybrid screening and immunoprecipitation assays, and the binding region was mapped to central regions in both proteins (Chiang and Roeder, 1995; Gegonne et al., 2001). The putative interaction region in TAF1 included a larger part of a proposed, putative histone acetyl transferase (HAT) domain and a further region that is thought to interact with the large subunit of human TFIIF, RAP74 (RAP74 Interacting Domain RAPiD) as well (Ruppert and Tjian, 1995). The interaction between TAF1 and TAF7 is thought to play a vital role in the transition from a static preinitiation complex including Pol II; where Pol II dissociates from the PIC and commences elongation (Gegonne et al., 2013). Trans- and auto-phosphorylation events involving the proposed kinase activity of TAF1 were implicated in this process, and TAF7 was suggested to inhibit the putative TAF1 HAT activity by binding to TAF1 (Gegonne et al., 2013).

Results from the laboratory of our collaborator, Dr. Christophe Romier (IGBMC Straßbourg) indicated that the interaction between TAF1 and TAF7 may consist of more than one potentially autonomous interaction interfaces. Romier mapped a stable interaction interface within TAF7 encompassing residues 126 to 202. This region of TAF7 interacted with a region encompassing residues 1157 to 1207 of TAF1, which is a part of the RAPiD domain. In these studies, it appeared that these rather small regions already stabilized a very tight and high affinity interaction between TAF1 and TAF7. This interaction on its own however would not explain the suggested inhibition of the acetyl transferase activity of TAF1 by TAF7, as the putative TAF1 HAT domain was not included.

3.1.1. Analysis of human TAF1/TAF7 complexes

In order to obtain insight into the interaction between TAF1 and TAF7, we produced and purified human TAF1/TAF7 complexes in insect cells using the MultiBac system to obtain samples with high purity and homogeneity that could be suitable for crystallization trials and

crystal structure determination by X-ray crystallography. We cloned a fragment of TAF1 encompassing amino acid residues 578 to 1210 into the pIDC Donor plasmid of the MultiBac system. This region encompasses the putative HAT domain of TAF1 and also the high affinity interaction region identified by Dr. Romier. Further, we cloned a fragment of human TAF7 encompassing amino acid residues 1 to 193 into the pFL Acceptor plasmid, comprising the predicted TAF1 interaction region of TAF7 including the short segment that was identified by Dr. Romier in the autonomous interaction region. We also prepared a second co-expression construct by cloning a shorter version of TAF1 encompassing residues 578 to 1106. $TAF1^{578\text{-}1106}$ was chosen based on multispecies conservation and secondary structure predictions (Appendix and Supplements, Chapter 5.4. & 5.5.). Although $TAF1^{578\text{-}1106}$ is missing ~100 C-terminal residues of proposed interaction region, it was designed to facilitate crystallization, based on the grounds that a long non-conserved ~50 AA stretch exists in $TAF1^{572\text{-}1206}$, which is predicted to be partially unstructured and could be highly flexible, thus impeding crystallization propensity. The $TAF1^{578\text{-}1106}$ construct was likewise cloned in pIDC. TAF1 constructs did not contain any affinity purification tags, while $TAF7^{1\text{-}193}$ was designed with a TEV cleavable N-terminal 10xHis-tag for metal affinity purification (IMAC). TAF1 and TAF7 constructs were fused by Cre-LoxP recombination *in vitro* and integrated by Tn7 transposition into the EMBacY virus contains a YFP reporter gene to monitor virus performance (Bieniossek et al., 2012). Complexes were expressed in insect cells following our published protocols.

$TAF1^{578\text{-}1210}$/$TAF7^{1\text{-}193}$ and $TAF1^{578\text{-}1106}$/$TAF7^{1\text{-}193}$ complexes were purified to homogeneity by using IMAC and SEC, as described in Methods (Chapter 2.2.5.1). His-tag was cleaved using TEV protease after IMAC. As shown in Figure 10B & C, purified complexes eluted as a single peak on a Superdex S200 10/300 column at elution volumes expected for the heterodimers, and no aggregation was observed. SDS-PAGE analysis confirmed that the protein complexes are highly pure and suitable for crystallization trials. Main peak fractions of both complexes were screened for crystal formation via nanoliter

high-throughput crystallization trials at the high-throughput crystallization facility (HTX) laboratory (EMBL Grenoble). In total, close to 900 crystallization conditions were screened for each complex covering a broad crystallization space.

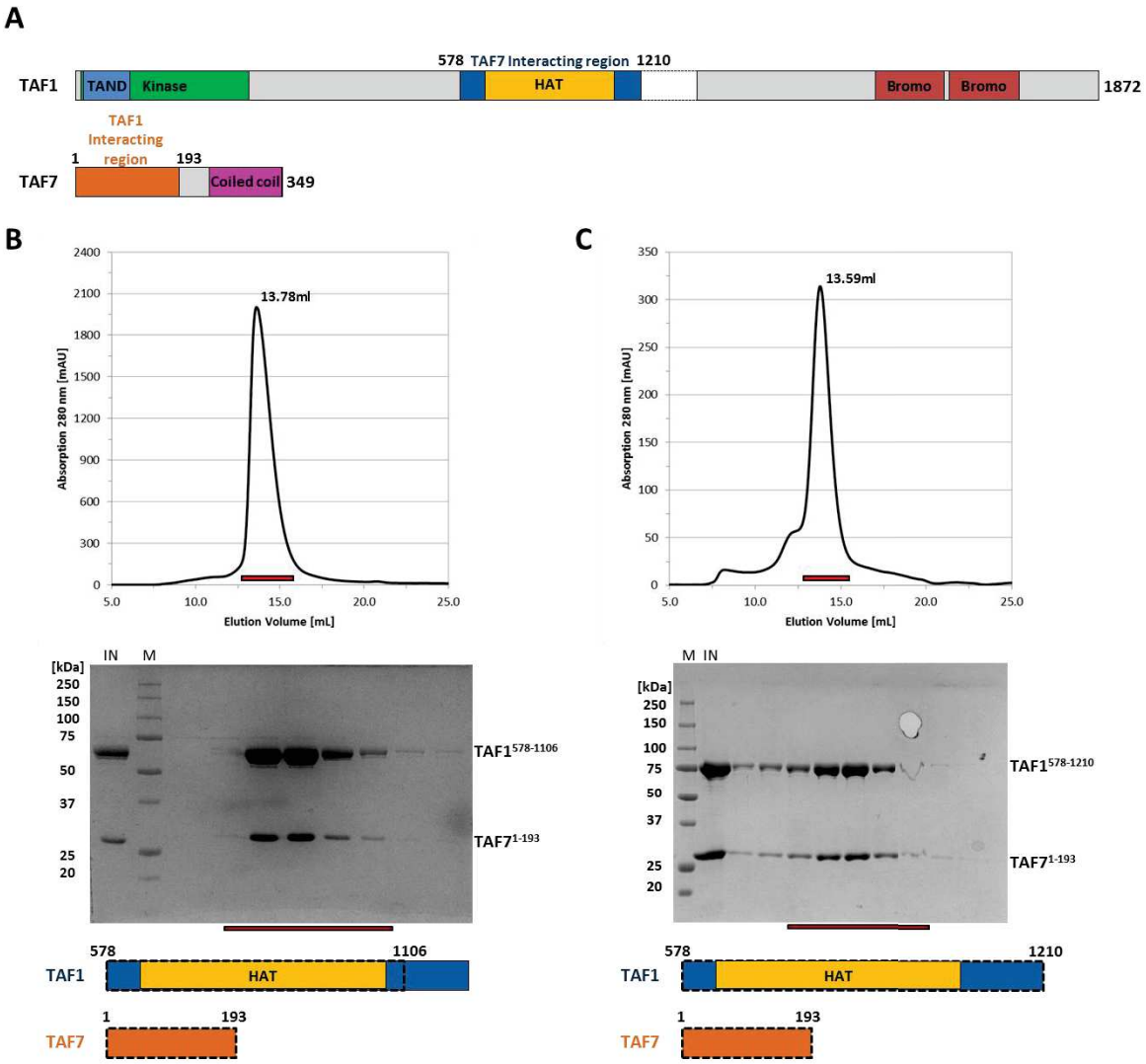


Figure 10: Interaction domains of TAF1/TAF7 complex.

(A) Domain representation of TAF1 and TAF7 along with interaction regions. **(B)** & **(C)** Final size exclusion chromatogram on Superdex S200 10/300 column and SDS-PAGE analysis of $TAF1^{578-1106}$/$TAF7^{1-193}$complex (B) and $TAF1^{578-1210}$/$TAF7^{1-193}$ complex (C), indicating nicely purified complexes.

We did not obtain crystals for the larger TAF1$^{578\text{-}1210}$/TAF7$^{1\text{-}193}$ complex, indicating that the large unstructured internal loop may indeed impede crystallization. SDS-PAGE analysis showed that TAF1$^{578\text{-}1210}$ in the TAF1$^{578\text{-}1210}$/TAF7$^{1\text{-}193}$ complex was progressively degrading by residual protease activity, even at 4°C. Interestingly, a band of similar size to TAF1$^{578\text{-}1106}$ accumulated in the degradation (data not shown), probably due to proteolytic cleavage in the flexible loop. In marked contrast, many crystallization hits were obtained for TAF1$^{578\text{-}1106}$/TAF7$^{1\text{-}193}$ complex (Figure 11B), which were manually reproduced and optimized in hanging- and sitting-drop 24-well plate setups as described in Methods (Chapter 2.2.5.2) to obtain rectangular shaped crystals (Figure 11C). A variety of cryo-protectants were screened for better diffraction, as these crystals diffracted differently with different cryo-protectants. Data collected from these crystals with 20% glycerol as cryo-protectant was anisotropic while, crystals from the same crystallization drop diffracted isotropically with 25% EG (Ethylene Glycol) as cryo-protectant. Presence of anisotropy was checked using Diffraction Anisotropy Server (http://services.mbi.ucla.edu/anisoscale/). Finally these crystals were used, with 25% EG as cryo-protectant, to collect 5 different native x-ray diffraction data sets up to 1.75 Å resolution at ID29 with a DECTRIS PILATUS 6M pixel detector in shutter-less collection mode. A representative diffraction images is shown in Figure 11D. The crystals belong to primitive orthorhombic space group $P2_12_12_1$. All these data sets were merged and processed as described in Methods (Chapter 2.2.5.2). Matthew's coefficient calculations suggested for the presence of one copy of complex in the asymmetric unit with high probability, corresponding to a solvent content of ~50 % (see Appendix and Supplements, Figure 31 & Table 12).

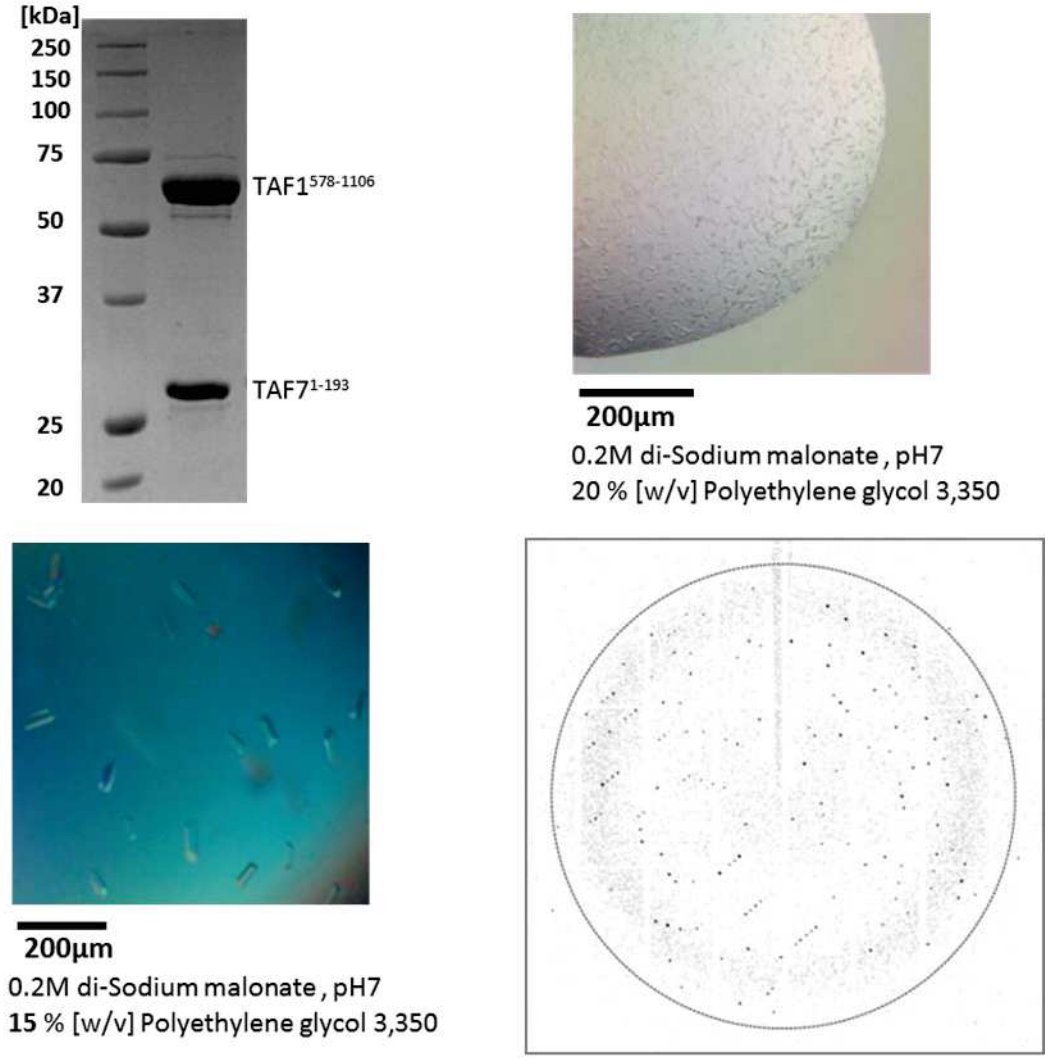


Figure 11: Crystallization and data collection of TAF1$^{578-1106}$/TAF7$^{1-193}$.

(A) Purified TAF1$^{578-1106}$/TAF7$^{1-193}$ used for crystallization. **(B)** One of the initial crystallization conditions and crystals from 96 well plate sitting drop. **(C)** Final optimized crystals and crystallization condition in 24 well hanging drop plate. **(D)** Diffraction image collected at ESRF beamline ID29.

Meanwhile, when we were finishing these studies, a structure of TAF1/TAF7 from *S. cerevisiae* (yeast) was published (Bhattacharya et al., 2014). This structure (PDB ID 4OY2) showed an intricate organization, comprising intertwined β-barrels and two arm domains. The TAF1/TAF7 interaction interface can be divided in two parts first an N-terminal interaction region within intertwined β-barrels and another C-terminal interaction region within right arm. The TAF1/TAF7 interaction regions corresponding to the domains in this

crystal structure are conserved between yeast and human, however, there are significant differences, specifically the insertion of the ~50 AA partially unstructured region which is not present in yeast TAF1 (see Appendix and Supplements, Chapter 5.1.

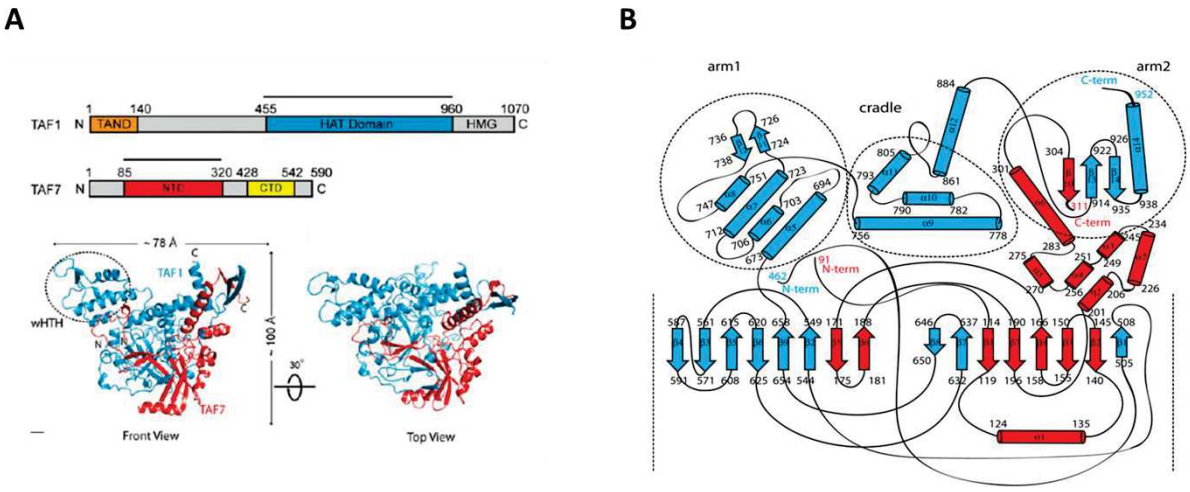


Figure 12: Crystal structure of *S. cerevisiae* (Yeast) TAF1/TAF7 complex.

(A) Domain representation of *S. cerevisiae* TAF1/TAF7 indicating the domains present in crystal structure and ribbon diagram representation of the crystal structure showing front and top view (PDB ID 4OY2). **(B)** Cartoon representation of the structure showing β-barrel and right and left arm. Taken from (Bhattacharya et al., 2014)

We utilized the yeast crystal coordinates to solve our structure by molecular replacement (MR) as described in Methods (Chapter 2.2.5.2). The initial model was manually adjusted and further improved by repetitive rounds of manual model building and refinement in COOT and PHENIX as described in Methods (Chapter 2.2.5.2). Final electron density map shows clear main chain and side chain density for almost all residues (Figure 13B). Final data collection and refinement statistics are shown in Table 10. The final model showed good stereochemistry and had similar quality as of many structures of similar resolution according to structure validation by MOLPROBITY server (see Appendix and Supplements, Figure 32)

The final crystal structure is shown in Figure 13A in various views. The structure comprises, in addition to protein, 541 water molecules, residues 588-1090 of TAF1 and residues 10-154 of TAF7. Electron density for first 9 N-terminal and last 16 C-terminal residues as well as residues 857-858, 909-921, 973-1053 of TAF1 and First 9 N-terminal and last 29 C-terminal residues as well as residues 107-123 of TAF7 was absent and these were not modelled in the final structure. These AA are predicted to be partially unstructured (see Appendix and Supplements, Chapter 5.4. & Figure 30). The C-terminal part of TAF7 which is present in the crystal is not visible in the structure, consistent with our finding that it is unstructured in the absence of its interaction partner $TAF1^{1157-1207}$ as confirmed by NMR (see also Chapter 3.1.2).

Table 10: Data collection and refinement statistics.

Data collection	
Wavelength (Å)	0.97625
Space group	$P2_12_12_1$
Cell constants	a=83.32 Å, b=95.44 Å, c=101.91 Å, α=β=γ= 90°
Resolution (Å)*	47.74-1.75 (1.85-1.75)
Unique reflections*	82479 (12500)
Completeness (%)*	100 (100)
Redundancy*	23.08 (22.91)
*I/σI**	12.36 (1.57)
R_{meas} (%)*	18.8 (>100)
$CC_{½}$ (%)*	99.9 (53.3)
Wilson B factor (Å^2)	26.8
Refinement	
R_{work}/R_{free} (%)	0.1630 /0.1844
Protein residues	533
No. of atoms	
Protein	8689
Hydrogen	4364
Water	541
B factor (Å^2)	
Protein	42.76
water	42.42
All	42.74
All (No hydrogen)	38.3
RMS deviation	
Bond angles (Å)	0.0073
Bond angles (°)	1.072
Ramachandran	
Favoured (%)	97.9
Allowed (%)	2.11
Outliers (%)	0
Clashscore	1.5

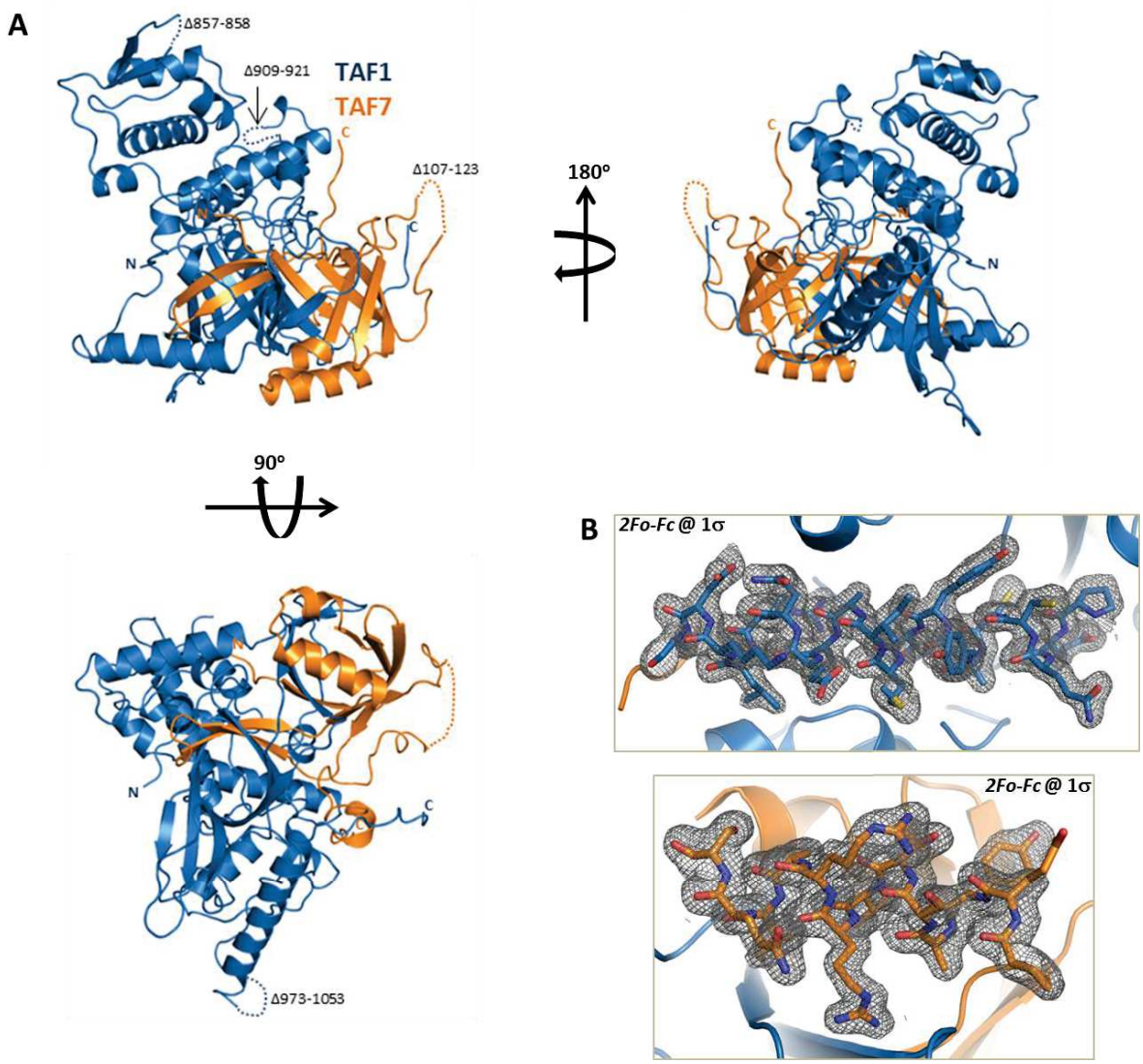


Figure 13: Crystal structure of TAF1$^{578-1106}$/TAF7$^{1-193}$.

(A) Final refined structure at resolution 1.75Å is shown at various angles. N and C terminus for both proteins are labelled 'N' and 'C' also missing loops which are not visible in crystal structure are shown with dotted lines. **(B)** Refined electron density map of TAF1$^{578-1106}$/TAF7$^{1-193}$ for one helix of TAF1 (top) and for one helix of TAF7 (bottom). All electron density maps are 2Fo-Fc, shown at contour level 1σ.

Our structure of human TAF1/TAF7 contains two intertwined β-barrels, where both TAF1 and TAF7 donate β-strands (Figure 14C). In one β-barrel, 5 β-strands were contributed by TAF1 and another 2 by TAF7, while in another β-barrel, 5 β-strands were contributed by TAF7 and another 3 by TAF1. Interestingly, the intertwined barrel resembles closely the

structure stabilizing the human TFIIF heterodimer (Gaiser et al., 2000). TAF1/TAF7 was superimposed on TFIIF (PDB ID 1F3U) with a root-mean-square deviation (RMSD) of 3.4 Å out of 271 aligned Cα atoms (Figure 14D). Second structure feature present in TAF1 is a winged helix domain consisting of two β-strands preceded by three α-helices (Figure 14A), which was structurally similar to DNA-binding winged helix domain of RAP30 subunit of human TFIIF (Groft et al., 1998). Presence of these structural domains suggests for a functional similarity of TAF1/TAF7 with TFIIF. TAF1/TAF7 was superimposed with winged helix of RAP30 (PDB ID 2BBY) with RMSD 1.9 Å out of 69 aligned Cα atoms (Figure 14E). Other secondary structures included three α-helices of TAF7 and 10 α-helices of TAF1. Additionally, Three TAF1 residues S1084, T1085 and T1087 were found to have phosphorylations, which are shown with refined electron density in Figure 14B. Physical significance of these phosphorylation sites still needs to be determined. The structure has overall dimensions of about ~65 Å x ~65 Å x ~70 Å. Total solvent accessible surface that is buried between two proteins was ~8394 $Å^2$, which suggests that the complex is highly stable.

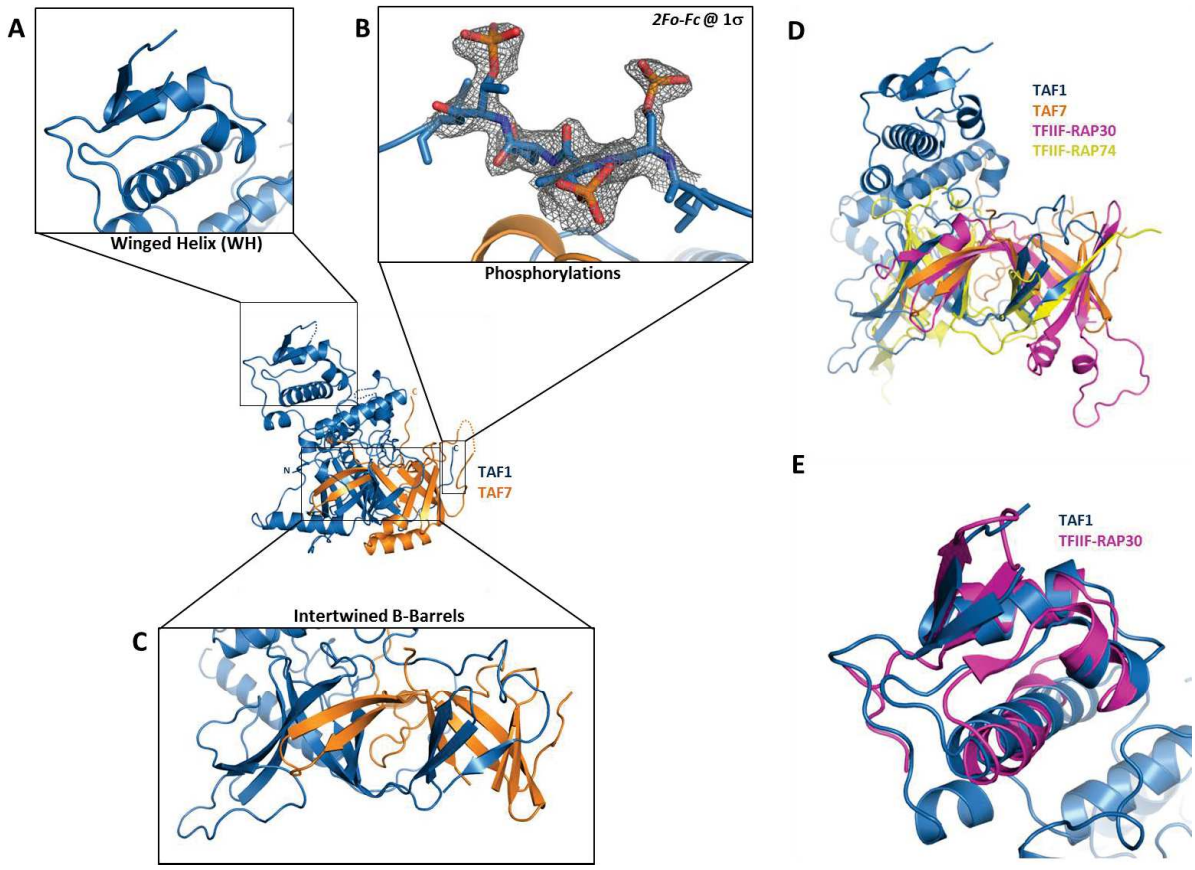


Figure 14: Structural features of TAF1$^{578-1106}$/TAF7$^{1-193}$.

(A) Winged helix domain of TAF1. **(B)** Refined electron density for phosphorylated TAF1 residues S1084, T1085, T1087. Functional significance of this phosphorylation is yet to be determined. Electron density map is 2Fo-Fc, shown at contour level 1σ. **(C)** Intertwined β-barrel of TAF1/TAF7, where β-strand comes from both TAF1 and TAF7. **(D)** Superposition of TAF1/TAF7 β-barrel with that of human TFIIF (RMSD=3.4 Å) **(E)** Superposition of TAF1 winged helix domain with that of RAP30 subunit of human TFIIF (RMSD=1.9 Å).

A comparison of our human TAF1/TAF7 structure with the yeast TAF1/TAF7 crystal structure (3Å resolution) is shown in Figure 15. Both structures superimpose with a root-mean square deviation (RMSD) of 2.0 Å. Major domains of TAF1/TAF7 are conserved between human and yeast (Figure 15A). In addition, there are yeast specific loops and α-helices which are absent from human at the amino acid level (see Appendix and Supplements, Chapter 5.1. & 5.2.). Similarly one α-helix is present only in human, which is

partially absent from yeast at amino acid level. TAF1 winged helix domain was also conserved but was moved towards the β-barrel by ~ 12-15 Å as shown in Figure 15B. A possible explanation for this might come from the structural hindrance created by yeast specific loop, which is absent in human. Also the pocket shown to be binding with histone tail peptide containing H3K27me3 marks in yeast was also found to be conserved in human. The TAF1 residues shown to be involved in this interaction in yeast (F480, F493 and H494) were conserved in human (F602, F615, H616), while TAF7 residue (F186) was not conserved in human (A89) (Figure 15C).

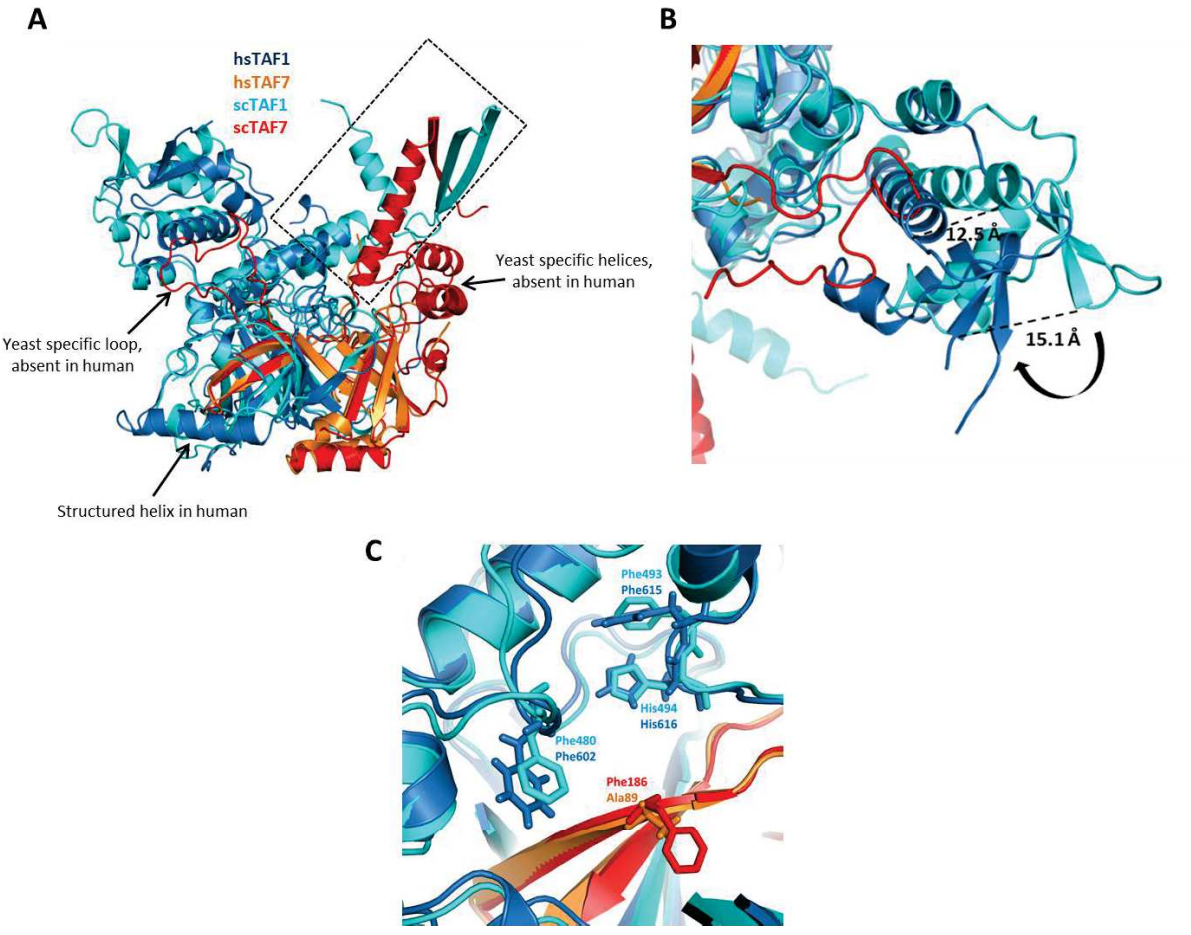


Figure 15: Comparison of Human TAF1$^{578-1106}$/TAF7$^{1-193}$ crystal structure with Yeast TAF1/TAF7.

(A) Superposition of structures from human and yeast (RMSD=2.0 Å) indicating yeast specific loop and α-helices as well as one human specific α-helix. C terminal domain is missing and not visible in human structure (TAF1 C terminal domain is not present in the construct), which is shown within the dotted box. **(B)** Movement of winged helix domain in human compared to yeast indicates the possible structural hindrance of a yeast specific loop with this domain, if present in human. Movement distances between corresponding residues are ~12-15 Å (shown with dotted lines). **(C)** Conservation of histone tail binding pocket in human.

Overall human TAF1/TAF7 crystal structure is adopting a similar shape as TAF1/TAF7 from yeast, with additional human specific features, however, the definition of the human structure is superior due to higher resolution (1.75 Å) as compared to yeast (3.0 Å). Furthermore, in our structure, we can see three phosphorylation sites which are not observed in the yeast complex.

Confirmation of our findings came at this stage from a crystal structure of a partial human TAF1/TAF7 complex, determined at 2.3 Å resolution (PDB ID 4RGW) (Wang et al., 2014), which showed mainly the intertwined β-barrel parts which superimposed reasonably on our coordinates (RMSD=0.199 Å). Interestingly, biochemical experiments in this study demonstrated that TAF1/TAF7 interacted with DNA (Wang et al., 2014).

3.1.2. NMR analysis and structure determination of TAF1$^{1157-1207}$/TAF7$^{153-190}$

In our crystal structure of human TAF1/TAF7, and also in the human structure determined by Wang and coworkers, the additional TAF1/TAF7 interaction region originally defined by our collaborator Dr. Romier, was absent (this was present in the yeast TAF1/TAF7 structure determined by Bhattacharya and coworkers). In contrast, the longer construct which would have contained also this motif, did not crystallize, possibly due to the unconserved flexible loop present in human TAF1. We had previously established, by biochemical experiments that these partial interaction domains have a high affinity interaction (Kd= ~100-200 nM), in

spite of its small size. With the aim to structurally characterize the complete human TAF1/TAF7 interaction including this C-terminal interaction region of human TAF1/TAF7, we adopted an integrated approach. Already, a number of experiments were performed including interaction studies by SEC between TAF7 constructs and TAF1$^{1157-1207}$, ITC and AUC experiments to determine binding constants and stoichiometry by Matthias Haffke, a fellow PhD student in our laboratory, with the aim to obtain crystals of the small high-affinity interaction motif, which did not yield any crystals. The size of this small high-affinity interaction domain is suitable for structure determination by NMR. Therefore, we prepared corresponding complexes for NMR analysis in collaboration with Dr. Martin Blackledge (IBS Grenoble, France).

To obtain samples of high purity and homogeneity for NMR studies, corresponding TAF1 and TAF7 constructs were cloned in pMAL and pET28(a) plasmids, respectively, for expression in *E.coli*. Residues 1157-1207 of TAF1 were cloned with a TEV cleavable MBP-tag at the N-terminus and with a non-cleavable 6xHis-tag at the C-terminus for affinity purification, while residue 126-202 of TAF7 were cloned with a non-cleavable N-terminal 10xHis-tag for affinity purification. MBP_TAF1$^{1157-1207}$ and TAF7$^{126-202}$ were purified separately to homogeneity using IMAC and SEC, as described in the Methods section (Chapter 2.2.5.3). As shown in Figure 16A & B, both purified proteins eluted as a major single peak on a Superdex S200 16/60 column (MBP_TAF1$^{1157-1207}$) and Superdex S200 10/300 column (TAF7$^{126-202}$) at elution volumes expected for monomeric proteins. There were some soluble aggregates for both proteins, which eluted in the void volume and could be separated from the main peak. SDS-PAGE analysis shows that the proteins are highly pure. Main peak fractions of both were checked for the presence of and found to be devoid of any more soluble aggregates; by reinjecting it on the Superdex S200 16/60 column (data not shown). ^{15}N and/or ^{13}C labelling for both proteins were carried out as described in Methods (Chapter 2.2.4.4 & 2.2.5.3). Purified proteins were used for collecting a variety of NMR spectra in collaboration with Dr. Malene Ringkjøbing Jensen in the Blackledge laboratory.

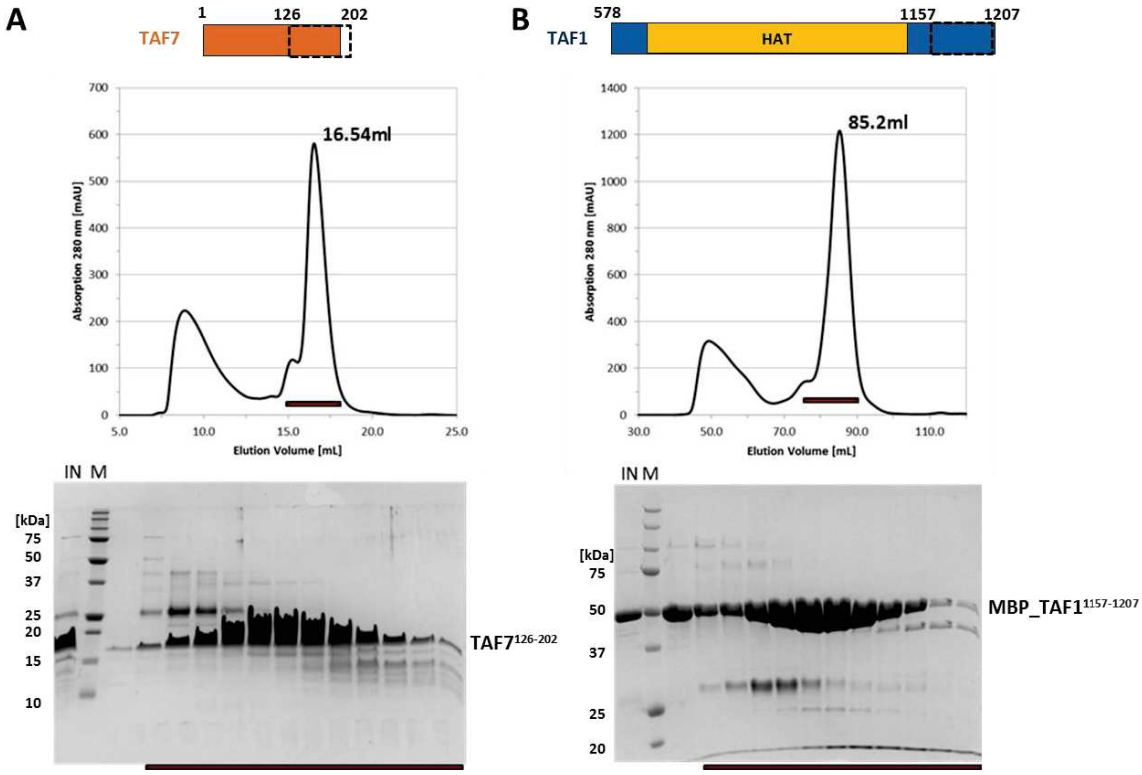


Figure 16: Purification of TAF1$^{1157-1207}$ and TAF7$^{126-202}$.

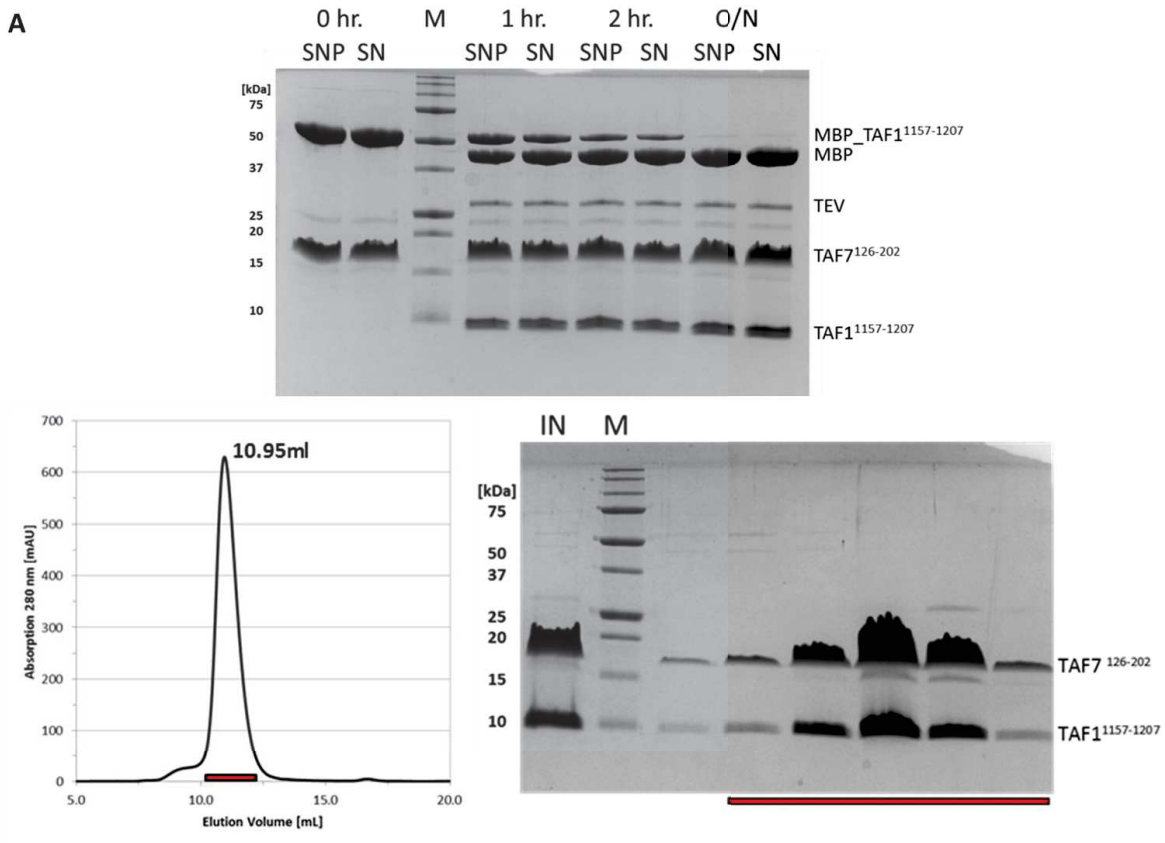
(A) & (B) Final size exclusion chromatogram and SDS-PAGE analysis of TAF7$^{126-202}$ on Superdex S200 10/300 column (A) and of MBP_TAF1$^{1157-1207}$ on Superdex S200 16/60 column (B), indicating nicely purified proteins (There were some minor impurities; which were removed in further steps of complex formation).

First, we recorded a two-dimensional ^{1}H-^{15}N HSQC NMR spectrum of TAF7$^{126-202}$. The spectrum shows a limited dispersion of the resonances in the ^{1}H dimension indicative of an intrinsically disordered protein (Figure 17 B). We then assigned the resonances using a set of triple resonance experiments allowing us to calculate the percentage of secondary structure from the experimental ^{13}C chemical shifts (Marsh et al., 2006). The chemical shifts show that the isolated TAF7$^{126-202}$ domain is devoid of significant secondary structure and essentially behaves as a random coil chain (Figure 18 C).

We also attempted to collect a ^{1}H-^{15}N HSQC spectrum of MBP_TAF1$^{1157-1207}$ however only very few resonances of low intensity could be detected (data not shown). We hypothesize that the TAF1 peptide may be transiently in contact with the surface of MBP thereby reducing the NMR signal intensities. We tried to remove the MBP tag for the NMR studies, but TAF1$^{1157-1207}$ precipitated when MBP was cleaved off (data not shown). In order to get TAF1$^{1157-1207}$ for NMR studies we also replaced the MBP tag with the much smaller GB1 and Trx tags. These tags have been shown to solubilize a protein similar to MBP and are suitable for NMR due to their smaller size (Huth et al., 1997; LaVallie et al., 2003). Apparently these tags did not make TAF1$^{1157-1207}$ sufficiently soluble (data not shown). Finally, we were able to cleave the MBP tag from TAF1$^{1157-1207}$, but only in presence of TAF7$^{126-202}$ (Figure 17A). Cleaved complex was further purified from cleaved MBP-tag, using IEX and SEC as described in Methods (Chapter 2.2.5.3). Purified TAF1$^{1157-1207}$/ TAF7$^{126-202}$ complex was eluted as a single peak on a Superdex S75 10/300 column, at elution volumes expected for a monomer of heterodimer (Figure 17A). Main peak fractions were used to collect ^{1}H-^{15}N HSQC spectra of ^{15}N-^{13}C labelled TAF7$^{126-202}$ alone and in complex with unlabelled TAF1$^{1157-1207}$ as well as for ^{15}N-^{13}C labelled TAF1$^{1157-1207}$ in complex with unlabelled TAF7$^{126-202}$ (Figure 17B). The dispersions of the resonances in the ^{1}H dimension in these spectra indicate that some residues of TAF7$^{126-202}$ get structured in the presence of TAF1$^{1157-1207}$ and that TAF1$^{1157-1207}$ is also mostly structured in the presence of TAF7$^{126-202}$.

By comparing the ^{1}H-^{15}N HSQC spectra of isolated TAF7$^{126-202}$ and TAF7$^{126-202}$ in complex with TAF1$^{1157-1207}$ (Figure 17), we could show that only a small part of TAF7$^{126-202}$ is involved in the complex formation corresponding to residues 153-190. Thus some peaks of the complex remain in the same position as in isolated TAF7$^{126-202}$ thereby corresponding to residues that are not directly involved in the interaction with TAF1$^{1157-1207}$. Using the assignment of the isolated TAF7$^{126-202}$, we identified that only residues 153-190 of TAF7 are

needed for interaction with TAF1$^{1157-1207}$ and a construct containing resides 153-190 of TAF7 was therefore used for further NMR studies and structure determination of the complex.



B

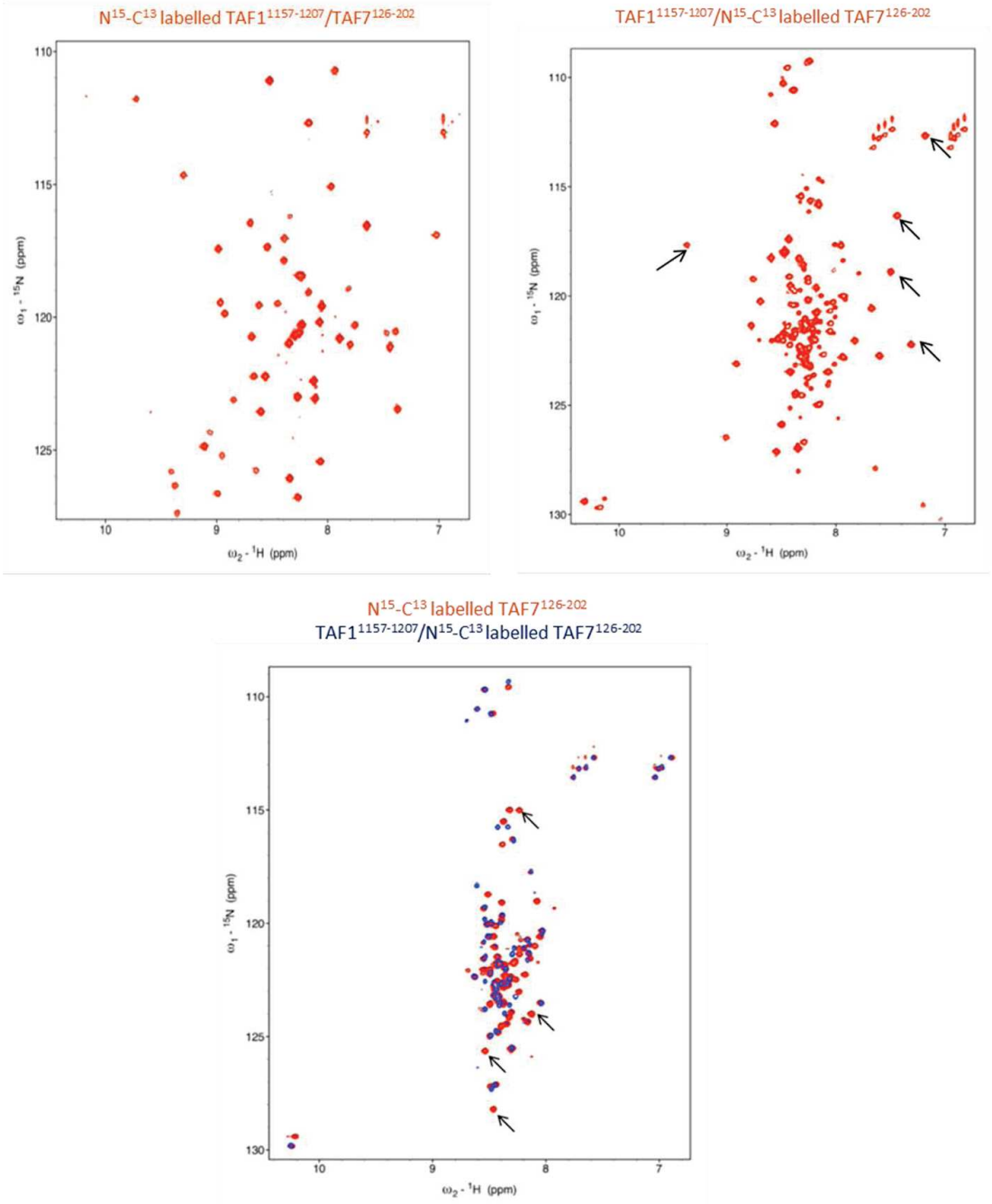


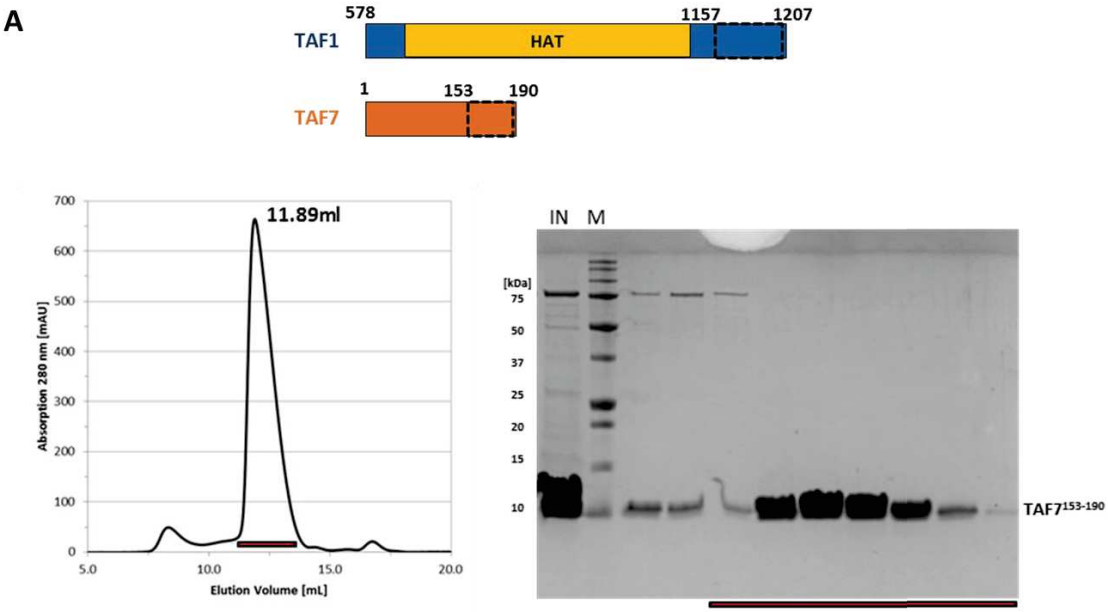
Figure 17: NMR analysis of TAF1$^{1157\text{-}1207}$/TAF7$^{126\text{-}202}$ complex.

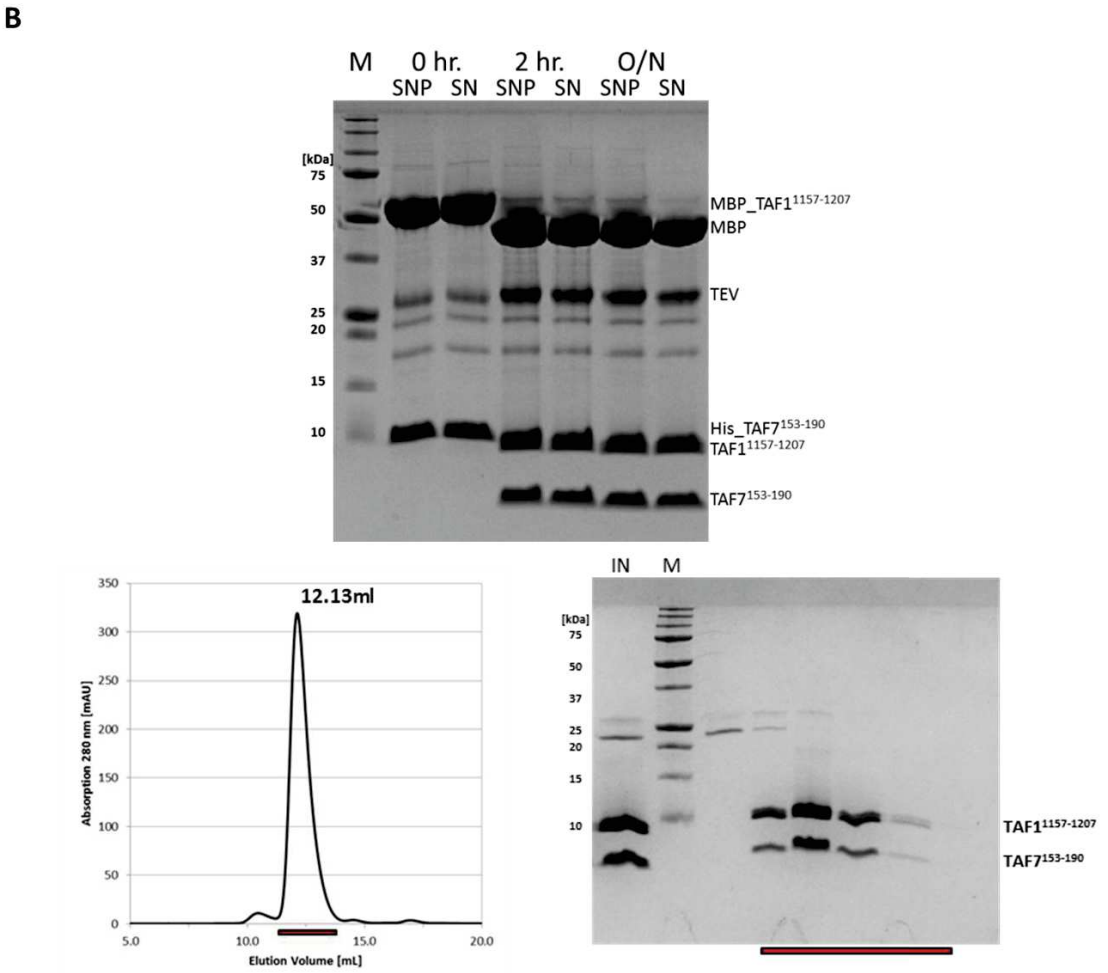
(A) Purification of TAF1$^{1157\text{-}1207}$/TAF7$^{126\text{-}202}$ complex without MBP tag. On the top SDS-PAGE analysis for TEV cleavage to remove MBP tag is shown. Below it final size exclusion chromatogram on Superdex S75 10/300 column and SDS-PAGE analysis of the cleaved complex after separation from MBP tag is shown. **(B)** ^{1}H-^{15}N HSQC spectra of differently labelled TAF1$^{1157\text{-}1207}$/TAF7$^{126\text{-}202}$ complexes. On top left

side ^{1}H-^{15}N HSQC spectrum for ^{15}N-^{13}C labelled TAF1$^{1157-1207}$/unlabelled TAF7$^{126-202}$ recorded at 25°C is shown. The dispersion in the ^{1}H dimension shows that TAF1$^{1157-1207}$ is folded in the complex. On the bottom superimposition of ^{1}H-^{15}N HSQC spectra of unlabeled TAF1$^{1157-1207}$/^{15}N-^{13}C labeled TAF7$^{126-202}$ recorded at 10°C (blue) and ^{15}N-^{13}C labeled TAF7$^{126-202}$ alone recorded at 10°C (orange) is shown . At 10°C, some peaks in TAF7$^{126-202}$ disappear upon formation of the complex with TAF1$^{1157-1207}$ (indicated by arrows), while other peaks remain in the same position as in isolated TAF7$^{126-202}$. The peaks that remain in the same position correspond to residues that are not directly involved in the interaction with TAF1$^{1157-1207}$. Using the assignment of the isolated TAF7$^{126-202}$, we identify that region of TAF7 which binds to TAF1$^{1157-1207}$, contains residues 153-190. On top right side ^{1}H-^{15}N HSQC spectrum for unlabelled TAF1$^{1157-1207}$/^{15}N-^{13}C labelled TAF7$^{126-202}$ recorded at 25°C is shown. At 25°C we increase the overall tumbling of the molecule in solution compared to the spectra recorded at 10°C, and we now see peaks that are significantly more dispersed in the ^{1}H dimension showing that a part of TAF7 folds into a three dimensional structure upon interaction with TAF1$^{1157-1207}$ (indicated by arrows).

Therefore, a novel construct of TAF7 containing residues 153-190 was cloned in pET28(a) plasmid with TEV cleavable N-terminal 10xHis-tag for affinity purification. TAF7$^{153-190}$ was purified to homogeneity as described in Methods (Chapter 2.2.5.3). A single peak was obtained for purified TAF7$^{153-190}$ on Superdex S75 10/300 column (Figure 18A). TAF7$^{153-190}$ was further complexed with TAF1$^{1157-1207}$, then MBP tag was cleaved from TAF1$^{1157-1207}$ (Figure 18B) and cleaved complex was further purified from MBP-tag, using IEX and SEC as described in Methods (Chapter 2.2.5.3). Purified TAF1$^{1157-1207}$/ TAF7$^{153-190}$ complex was eluted as a single peak on a Superdex S75 10/300 column (Figure 18B). To be noted here, the His-tag on TAF7$^{153-190}$ was also TEV cleavable here and got cleaved when MBP-tag was cleaved from TAF$^{11157-1207}$. The sample quality was checked by MS and the presence of monomer of heterodimer by SEC-MALLS (see Appendix and Supplements Figure 34). Two dimensional ^{1}H-^{15}N HSQC spectra for ^{15}N-^{13}C labelled TAF7$^{153-190}$ alone and in complex with unlabelled TAF1$^{1157-1207}$ as well as for ^{15}N-^{13}C labelled TAF1$^{1157-1207}$ in complex with unlabelled TAF7$^{153-190}$ were collected using main peak fractions from purified complex (Figure 18C). These spectra show that TAF1$^{1157-1207}$ is mostly structured in complex

with TAF7$^{153-190}$ and TAF7$^{153-190}$ also gets structured in presence of TAF1$^{1157-1207}$. HN, N, C_α, C' and C_β nuclei of backbone were also assigned for all the TAF1 and TAF7 residues as shown in Figure 18C. Side chains were assigned using three dimensional ^{1}H-^{13}C HCCH-TOCSY spectra. Secondary structure content calculated from C_α and C_β chemical shifts showed that TAF7$^{153-190}$ is unstructured and gets structured in presence of TAF$^{11157-1207}$ with one α-helix and one β-strand. TAF1$^{1157-1207}$, in presence of TAF7$^{153-190}$, also contains two β-strands and one α-helix (Figure 18C). Also, spectra for TAF1$^{1157-1207}$ and TAF7$^{153-190}$ are not overlapping, so both of them were labelled in the same sample to collect three-dimensional ^{1}H-^{13}C HSQC-NOESY and ^{1}H-^{15}N HSQC-NOESY spectra to get distance restraints. All these data were used to determine the NMR solution structure.


A
578
1157
1207
TAF1
HAT
1
153
190
TAF7
11.89ml
Absorption 280 nm [mAU]
Elution Volume [mL]
IN
M
[kDa]
75
50
37
25
20
15
10
TAF7153-190


B
M
0 hr.
2 hr.
O/N
SNP
SN
SNP
SN
SNP
SN
[kDa]
75
50
37
25
20
15
10
MBP_TAF11157-1207
MBP
TEV
His_TAF7153-190
TAF11157-1207
TAF7153-190
12.13ml
Absorption 280 nm [mAU]
Elution Volume [mL]
IN
M
[kDa]
75
50
37
25
20
15
10
TAF11157-1207
TAF7153-190

C

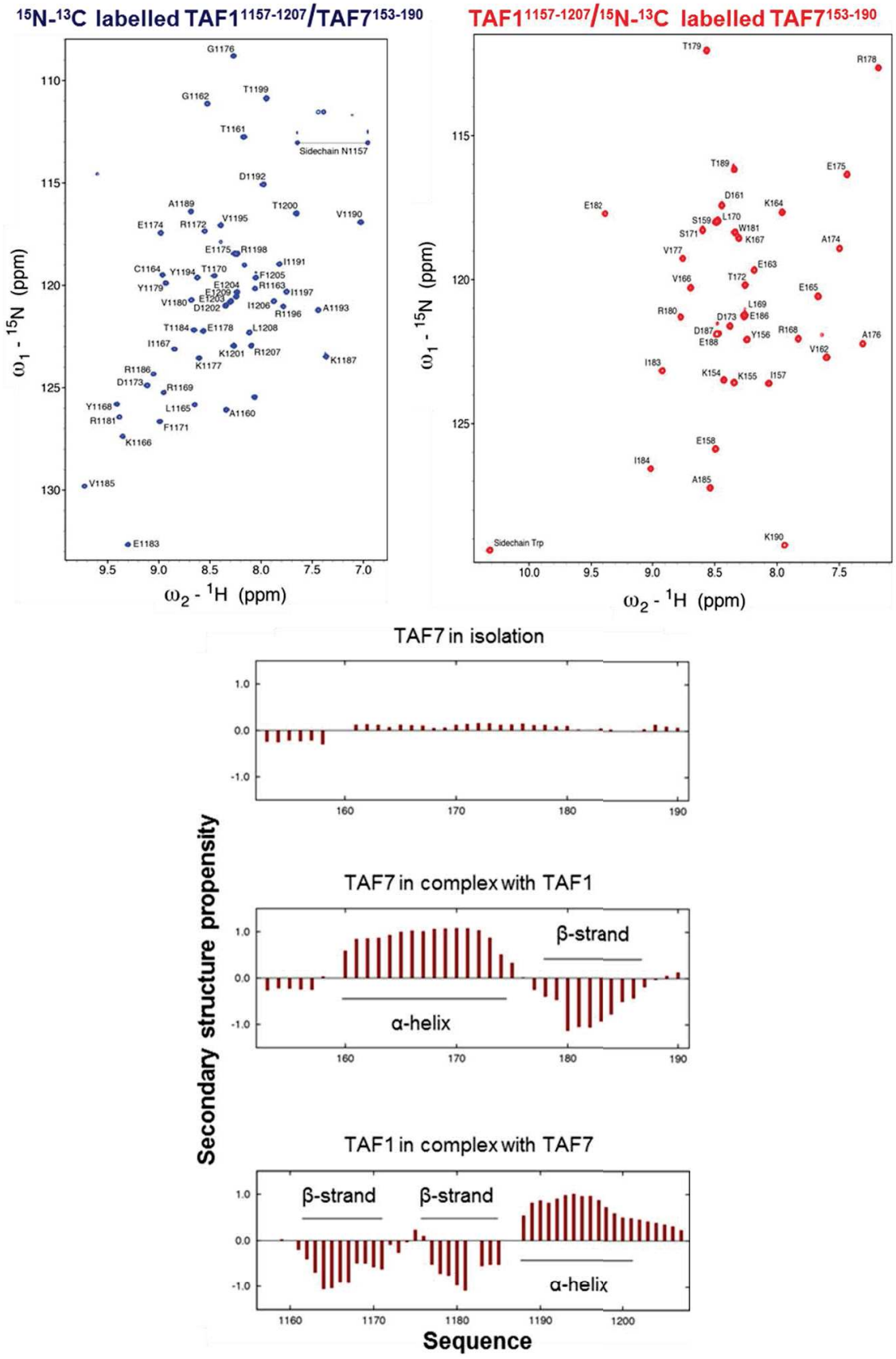

15N-13C labelled TAF1 1157-1207/TAF7 153-190
TAF1 1157-1207/15N-13C labelled TAF7 153-190
ω1 - 15N (ppm)
ω2 - 1H (ppm)
Sidechain N1157
Sidechain Trp
TAF7 in isolation
TAF7 in complex with TAF1
TAF1 in complex with TAF7
β-strand
α-helix
Secondary structure propensity
Sequence

Figure 18: NMR analysis of TAF1$^{1157-1207}$/TAF7$^{153-190}$ complex.

(A) Domain representation of TAF1 and TAF7 interaction region only indicating the position of TAF1$^{1157-1207}$ and TAF7$^{153-190}$. Final size exclusion chromatogram on Superdex S75 10/300 column and SDS-PAGE analysis of TAF7$^{153-190}$ is also shown. **(B)** Purification of TAF1$^{1157-1207}$/TAF7$^{153-190}$ complex without MBP tag. On the top SDS-PAGE analysis for TEV cleavage to remove MBP tag is shown. Below it final size exclusion chromatogram on Superdex S75 10/300 column and SDS-PAGE analysis of the cleaved complex after separation from MBP tag is shown. **(C)** NMR analysis of differently labelled TAF1$^{1157-1207}$/TAF7$^{153-190}$ complexes. ^{1}H-^{15}N HSQC spectrum for ^{15}N-^{13}C labelled TAF1$^{1157-1207}$/unlabelled TAF7$^{153-190}$ is shown in blue color indicating that TAF1$^{1157-1207}$ is mostly structured. ^{1}H-^{15}N HSQC spectrum for unlabelled TAF1$^{1157-1207}$/^{15}N-^{13}C labelled TAF7$^{153-190}$ is shown in red color indicating that TAF7$^{153-190}$ also have many structured residues. Label on each peak corresponds to the assignment of the backbone amide groups. On the bottom the secondary structure content of two proteins calculated from C_α and C_β chemical shifts is shown. The scale on the y-axis goes between -1 and 1. 1 corresponds to the formation of 100% α-helix, while -1 corresponds to the formation of 100% β-strand, indicating TAF7$^{153-190}$ is mostly unstructured in isolation and adopts a single α-helix followed by a β-strand in presence of TAF1$^{1157-1207}$, while TAF1$^{1157-1207}$ adopts two β-strands followed by one α-helix in the presence of TAF7$^{153-190}$.

NMR solution structure of TAF1$^{1157-1207}$/TAF7$^{153-190}$ is shown in Figure 19. Ensemble of 15 lowest energy structures, one of the best representative structure and a comparison with yeast TAF1 and TAF7 is shown. This structure was found to contain two β-strands and one α-helix of TAF1 and one α-helix and one β-strand of TAF7. Comparison with yeast structure shows that the structure is similar to structure of yeast right arm. Data collection and refinement statistics is shown in Table 11.

TAF1$^{1157-1207}$/TAF7$^{153-190}$ complex was also screened for crystallization as well at different concentrations at HTX lab, EMBL, Grenoble. Some crystals were obtained but they turned out to be salt (data not shown).

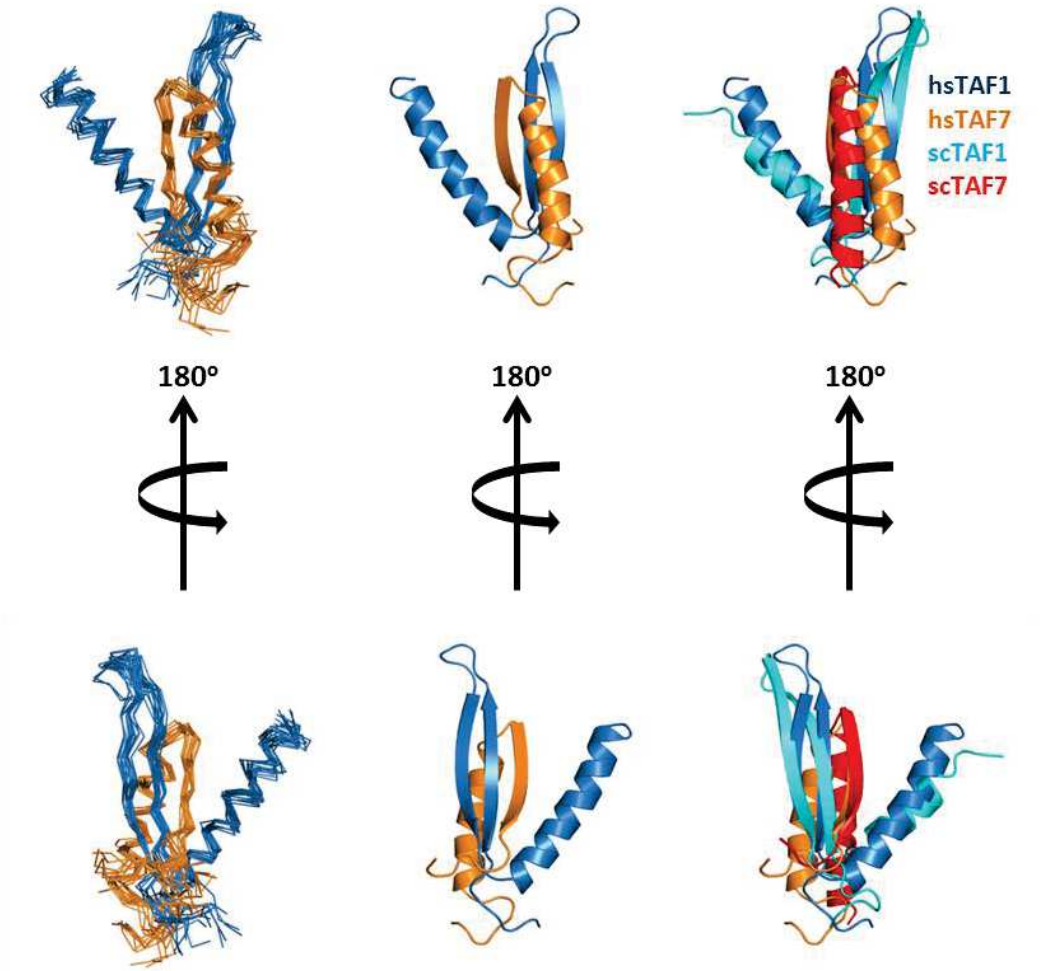


Figure 19: NMR structure of TAF1$^{1157-1207}$/TAF7$^{153-190}$ complex.

Ensemble of 15 lowest energy solution structures (left), one of the representative structure (middle) and comparisons with corresponding domain of yeast TAF1/TAF7 (PDB ID 4OY2) (right) are shown at two different angles, indicating a complex interaction between TAF1 and TAF7. Comparison indicates that the structural motif is conserved between human and yeast.

Table 11: NMR data collection and refinement statistics.

Resonance assignment completeness (%)	
Backbone H	100
Backbone non-H	92 (100% 1160-1209, 154-190)
Side-chain H	85
Side-chain non-H	90
Distance restraints	
Intraresidue (i − j = 0)	0
Sequential (|i − j| = 1)	204
Medium range (1 ≤ |i − j| ≤ 4)	100
Long range (|i − j| ≥ 5)	50
Total	594
Restraints per residue	6.6
Dihedral restraints	134
Phi angles	67
Psi angles	67
Validation Statistics	
Restraints statistics, average per structure	6.6
NOE distance restraint violations	
>0.1 Å	2±1
>0.5 Å	0
NOE energy per restraint (min/mean/max)	
NOE restraint weight (min/mean/max)	
Dihedral violations 1°–5°	4±1
Ramachandran plot quality (Procheck)	
Most favored regions (%)	84.3
Additional allowed regions (%)	12
Generously allowed regions (%)	2.4
Disallowed regions (%)	1.2

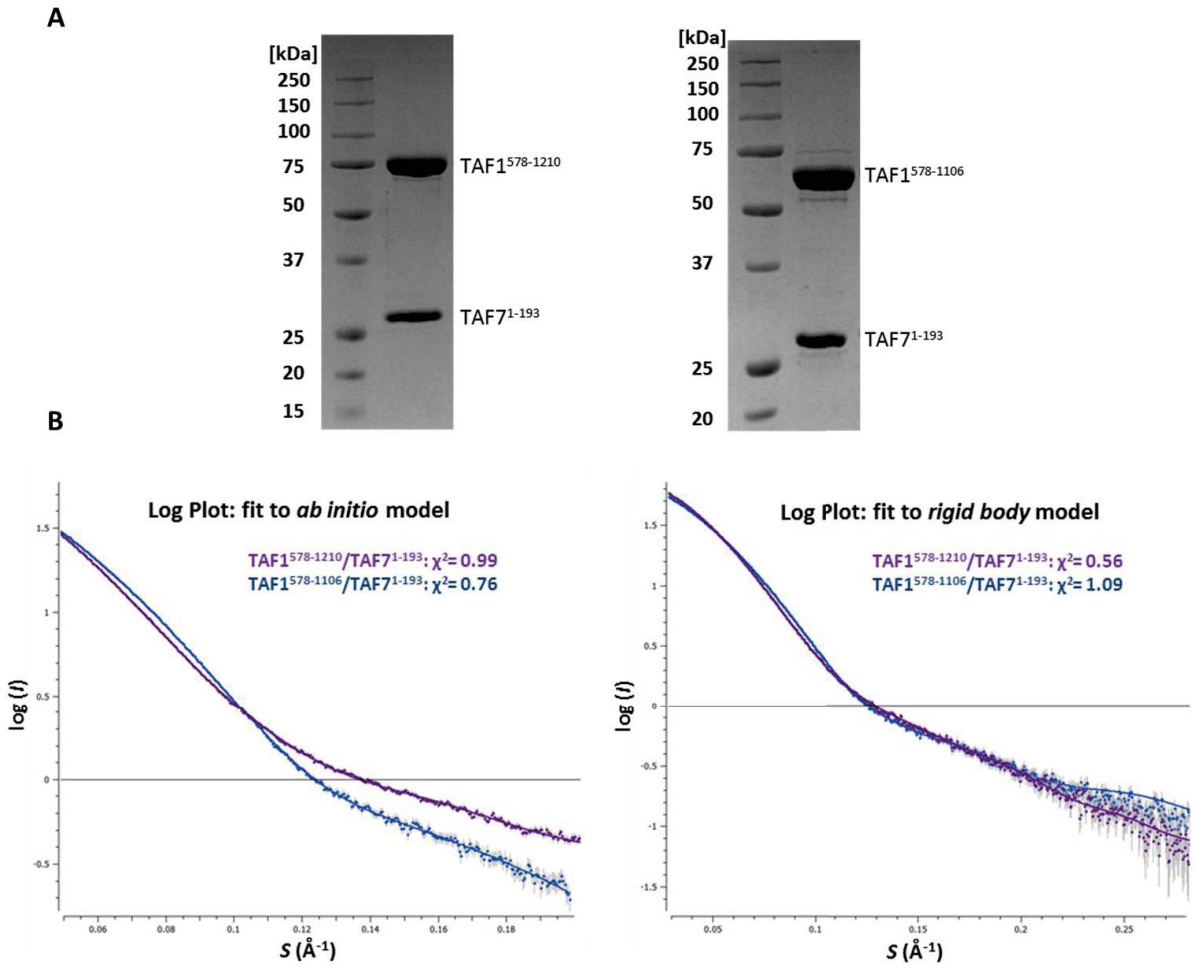
3.1.3. SAXS analysis of TAF1$^{578-1210}$/TAF7$^{1-193}$ complex and TAF1$^{578-1106}$/TAF7$^{1-193}$ complex

We were now in possession of atomic coordinates for the small high affinity TAF1/TAF7 interaction motif derived from NMR studies, and likewise of the large TAF1/TAF7 interaction domain containing the putative HAT domain of TAF1 from our X-ray

crystallographic analysis. In order to obtain complete structural information for the entire human TAF1/TAF7 interaction encompassing both interacting parts in solution, small angle X-ray scattering (SAXS) studies were performed. Our original construct, $TAF1^{578-1210}/TAF7^{1-193}$ complex, and also the crystallized $TAF1^{578-1106}/TAF7^{1-193}$ complex, were purified for SAXS as described in Methods (Chapter 2.2.5.1). Main peak fraction of freshly purified samples were used to collect static SAXS data at ESRF BioSAXS beamline BM29 with Pilatus 1M (with the help of Dr. Adam Round). Scattering curves were collected for different concentrations of samples with 10 frames for each concentration (see Appendix and Supplements, Table 13). These were averaged and subtracted for scattering from buffer, automatically with ATSAS software package, installed at beamline. Data was further analyzed as described in Methods (Chapter 2.2.5.5). Scatterings from different concentrations were then merged (constant Rg values) to decrease noise. Data analysis suggested that no aggregation was present, as shown by Guinier plot analysis. Analysis by Kratky plot indicated flexibility in both complexes (see Appendix and Supplements, Figure 33). Data quality was good as shown by different classical parameters used for validation of SAXS experiments (see Appendix and Supplements, Table 13). Rg calculated from Guinier analysis were 34.2 ± 0.05 Å and 33.4 ± 0.4 Å for $TAF1^{578-1210}/TAF7^{1-193}$ and $TAF1^{578-1106}/TAF7^{1-193}$ complexes respectively, while Dmax were 142.9 Å and 136.6 Å respectively.

Ab initio models for both the constructs were calculated with good χ^2 values of 0.99 and 0.76, respectively for $TAF1^{578-1210}/TAF7^{1-193}$ complex and $TAF1^{578-1106}/TAF7^{1-193}$ complex and with a nice fit to the data as described in Methods (Chapter 2.2.5.5). Rigid body models using crystal coordinates and the NMR structure were also generated as described in Methods (Chapter 2.2.5.5) to also model the missing loops. Best models had χ^2 values of 0.56 and 1.09, respectively for $TAF1^{578-1210}/TAF7^{1-193}$ complex and $TAF1^{578-1106}/TAF7^{1-193}$ complex and a nice fit to the data (Figure 20B). Different missing loops were incorporated between structured domains to compensate for the flexibility of the TAF1/TAF7 complex in solution, as indicated by SAXS, which lead to rearrangement of some of the α-helices. The

rigid body models were nicely fitting to the *ab initio* models (Figure 20C) and corroborated a structural reorganization when the TAF7$^{153-190}$ region gets structured in the presence of TAF1 (see chapter 3.1.2). The superimposition of both *ab initio* models with the rigid body model of TAF1$^{578-1210}$/TAF7$^{1-193}$ complex also showed extra volume in case of the larger TAF1$^{578-1210}$/TAF7$^{1-193}$ complex, where the NMR structure can be excellently fitted (Figure 20D). Our hybrid structure based on X-ray crystallography, NMR and SAXS thus provides structure of the complete human TAF1/TAF7 interaction.



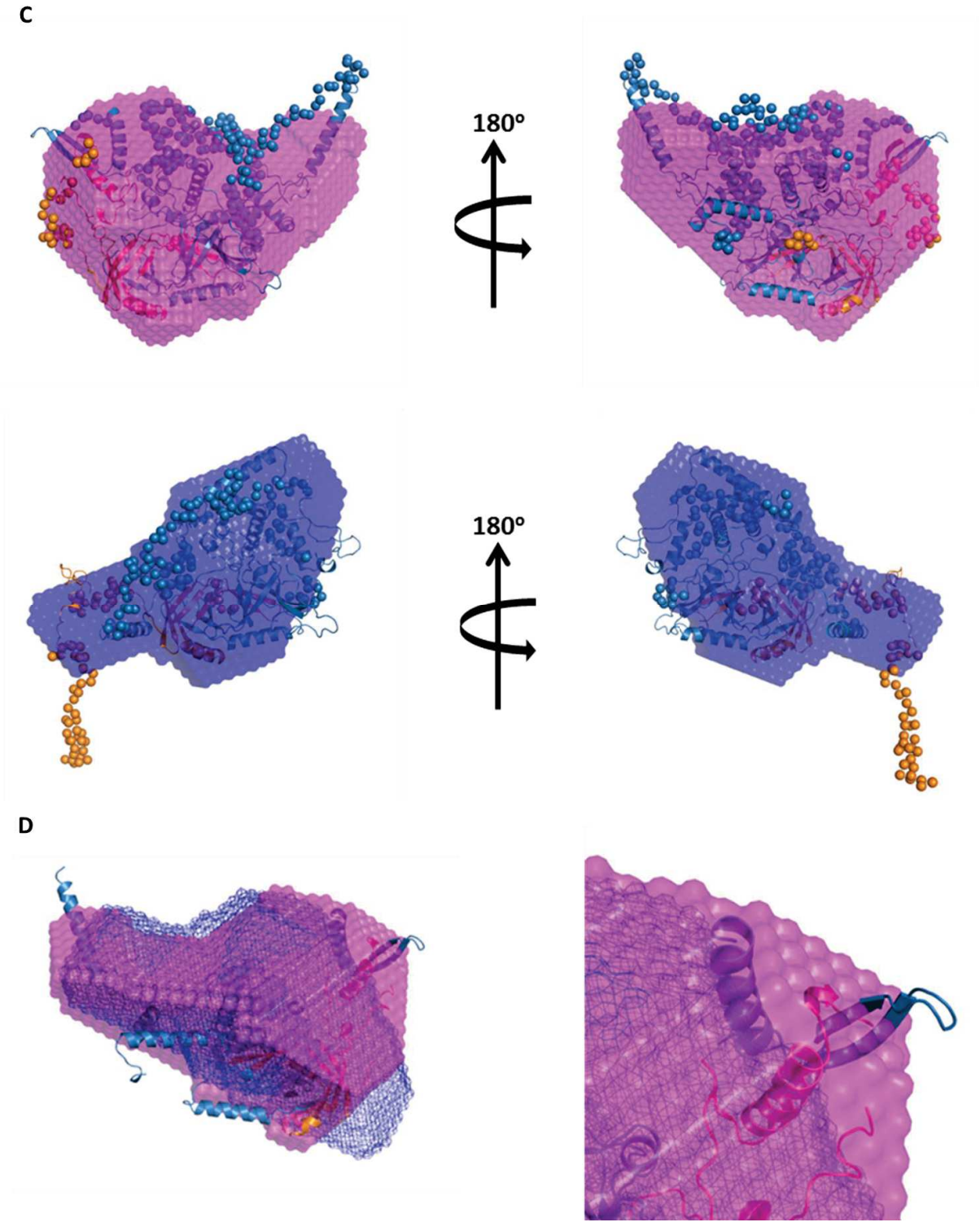

C
180°
180°
D

Figure 20: SAXS analysis of TAF1$^{578-1210}$/TAF7$^{1-193}$ complex and TAF1$^{578-1106}$/TAF7$^{1-193}$ complex.

(A) Purified TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ complexes used for SAXS analysis. **(B)** Log plot for fit of experimental scattering (shown as dots) to the calculated scattering from ab initio and rigid body models (shown as solid lines). TAF1$^{578-1210}$/TAF7$^{1-193}$ complex is shown in magenta color and TAF1$^{578-1106}$/TAF7$^{1-193}$ complex is shown in light blue color, as surface representation. Error bars are shown in grey. Corresponding χ^2 values for different models are also shown. **(C)** Overlay of *ab initio* and rigid body models for TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ complexes are shown here from two different angles. *Ab initio* models are shown in magenta and light blue color for TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ complexes respectively. TAF1 and TAF7 are shown respectively in sky blue and light orange colors. Residues present in crystal and NMR structures are shown with cartoon while the missing loops modelled with SAXS are shown with spheres. These overlays show that crystal and NMR structures are nicely fitting into the ab initio models. **(D)** Overlay of ab initio models for TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ complexes as wells as the crystal and NMR structures, indicating the NMR structure present in the extra density of TAF1$^{578-1210}$/TAF7$^{1-193}$. Please note that TAF1$^{578-1106}$/TAF7$^{1-193}$ ab initio model is shown with mesh representation here.

3.1.4. Interaction analysis by peptide array

Peptide array was performed for TAF7$^{153-190}$ with Alanine scan to find out the residues involved in interaction with TAF1$^{1157-1207}$. Analysis from peptide array showed that V162, V166, L169, L170, D173, T179, W181, I183 were involved in interaction with TAF1$^{1157-1207}$. Position of these residues in NMR structure of TAF1$^{1157-1207}$/TAF7$^{153-190}$ clearly shows that these residues are involved in interaction with the α-helix of TAF1 in C-terminal interaction region (Figure 21). CD spectra were also collected for TAF7$^{126-202}$ and TAF7$^{153-190}$ alone as well as along with TAF1$^{1157-1207}$, which also showed that TAF7$^{153-190}$ gets structured in the presence of MBP-TAF1$^{1157-1207}$, supporting the NMR results (data not shown). Different point mutant for all these residues were generated, which will be used for further mutational studies.

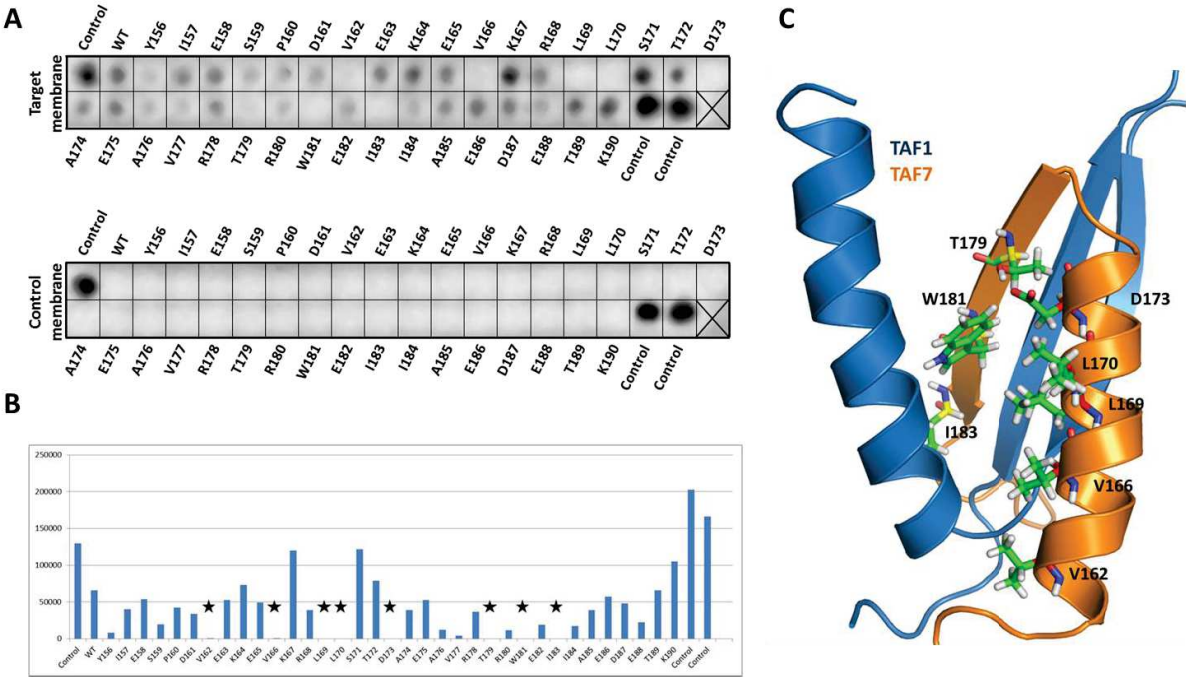


Figure 21: Peptide array (alanine scan) for TAF7$^{156\text{-}190}$.

(A) Peptide array membranes with immobilized TAF7$^{156\text{-}190}$ peptides, where different amino acid were mutated to alanine (one at a time), blotted against His-tag after treating with target or control protein are shown. Corresponding amino acid are shown next to each position. Control is for positive control (6xHis) and WT is for wild type TAF7$^{156\text{-}190}$ peptides. Last empty position is shown with a cross. Top membrane was reacted with target MBP_TAF1$^{1157\text{-}1207}$_His, while the bottom membrane was reacted with control His_MBP. **(B)** Normalized quantitative analysis of blotting at each position shows the major TAF7 residues involved in interaction with TAF1 in C-terminal interaction region. (C) Position of these TAF7 residues (shown as sticks) in NMR structure of TAF1$^{1157\text{-}1207}$/TAF7$^{153\text{-}190}$ confirms the involvement of these TAF7 residues in interaction with C-terminal interaction region of TAF1.

3.2. Characterization of a novel TAF11/TAF13/TBP complex

The TATA-box binding protein is a subunit of TFIID which can interact with the TATA box in the core-promoter. The crystal structure of TBP has been determined, and resembles a crescent shape with convex and concave surfaces. Promoter DNA binding by TBP is mediated by its concave surface. TBP is a dimer in the crystal, and was shown to exist as a dimer also in solution in the absence of DNA. The same concave surface which is implicated in DNA binding also serves as the homodimerization interface. The GTF TFIIA was shown to destabilize TBP dimers in the presence of DNA, and to stabilize the TBP/DNA interaction by forming a stable ternary TFIIA/TBP/DNA complex. Genetic and biochemical experiments lead to the suggestion that this TFIIA/TBP/DNA complex is further stabilized by TAF11 (Kraemer et al., 2001) (see also Chapter 1.3.2). These experiments were carried out before the prevalence of histone fold domains in TAFs was discovered. Later, it was shown that TAF11 heterodimerizes with TAF13, and the biochemical experiments that had been carried out only with TAF11 previously were repeated by substituting TAF11 with a preformed TAF11/TAF13 complex (Robinson et al., 2005). Altogether, the results from the experiments were interpreted to be consistent with the formation of a pentameric TAF11/TAF13/TFIIA/TBP/DNA complex. We originally set out to elucidate the structure of this pentameric complex. However, when we attempted to reproduce the published experiments, we found that the proposed pentameric complex was not formed in our hands. Instead, our experiments revealed a novel complex composed only of TAF11/TAF13/TBP, which seemingly had lost the competence to interact with DNA and TFIIA. Our experiments moreover showed that TAF11/TAF13 effectively competed with DNA for TBP binding. Furthermore, we could show that addition of TAF11/TAF13 to TBP results in monomeric TBP that is tightly bound by the TAF11/TAF13 complex. Together, our data is reminiscent of to the so-called TATA-box mimicry that was described for a domain in TAF1, TAF1-TAND (see Chapter 3.2.3). We characterized this novel TAF11/TAF13/TBP complex we discovered by using a variety of means in the present thesis.

3.2.1. TAF11/TAF13 dissociates DNA from preformed TFIIA/TBP/DNA complex

We analyzed the impact of TAF11/TAF13 on TFIIA/TBP/DNA complex formation by EMSA (our intention being to demonstrate reconstitution of the putative pentameric complex). For EMSA, human TAF11/TAF13, full-length human TBP (TBP-FL) as well as the conserved human TBP core (TBPc) were purified to homogeneity, as described in the Methods (Chapter 2.2.6). TBP has a highly conserved structure core and a divergent N-terminus. This conserved core has been shown to be fully functional (Cormack et al., 1991; Stevens, 1960) (See also Chapter 1.4.). Moreover, fully functional recombinant single-chain TFIIA (scTFIIA) was also expressed and purified. This scTFIIA is a single chain form of human TFIIA containing all three subunit polypeptides of the conserved TFIIA core in a single-chain format connected by linkers to facilitate production and purification. This scTFIIA has been shown to be completely active in Berger lab. Structure of scTFIIA was also determined in Berger lab; which is identical to unlinked TFIIA (unpublished, Isai Kandiah). TAF11/TAF13 complex was expressed in insect cells using our MultiBac system, while all other proteins for this study were expressed in *E.coli*. Final purified samples are shown in Figure 22A. Adenovirus major late promoter (AdMLP) dsDNA was prepared from synthetic oligonucleotides giving rise to 16 bp DNA oligo duplex containing a TATA box (sequence of coding strand: 5'ctgctataaaaggctg 3'). Quality of dsDNA samples was assessed by PAGE. Highly pure protein and dsDNA were used to attempt formation of the proposed pentameric complex, TAF11/TAF13/TFIIA/TBP/DNA.

EMSA was used to assay the formation of a stable complex (Figure 22B). First, the TBP/DNA complex was prepared, with TBP-FL or TBPc, respectively (Figure 22B). TBP/DNA complex is known to not produce a clean shift in the absence of TFIIA. Then, scTFIIA was added to form the trimeric TFIIA/TBP/DNA complex, resulting in a clean shift for both TBP-FL and TBPc, as expected. The amount of scTFIIA was titrated carefully to achieve this clean band-shift, thus forestalling addition of an excess amount of scTFIIA (data

not shown). Next, TAF11/TAF13 was added in increasing amounts to reconstitute the putative TAF11/TAF13/TFIIA/TBP/DNA complex.

Surprisingly and contrary to my expectations, a supershift of the TFIIA/TBP/DNA complex did not occur in these experiments, indicating that this putative pentameric complex was actually not formed (neither with TBP-FL, nor with the conserved TBP core). Instead, DNA was eliminated from the TFIIA/TBP/DNA complex upon addition of TAF11/TAF13, as evidenced by the increasing intensity of the free dsDNA band (Figure 22B). This observation was fully reproducible as confirmed by repetition of the experiments. Together with the absence of a supershift, this result effectively ruled out the formation of the proposed pentameric complex. Of note, TBP-FL and TBPc both produced identical results in EMSA. TBPc is the conserved C-terminal core of TBP devoid of the evolutionary divergent, unstructured N-terminal extension. TBPc has been shown to be fully functional as described earlier. Therefore, we continued our experiments with human TBPc and refer to this construct as ‘TBP’ hereafter.

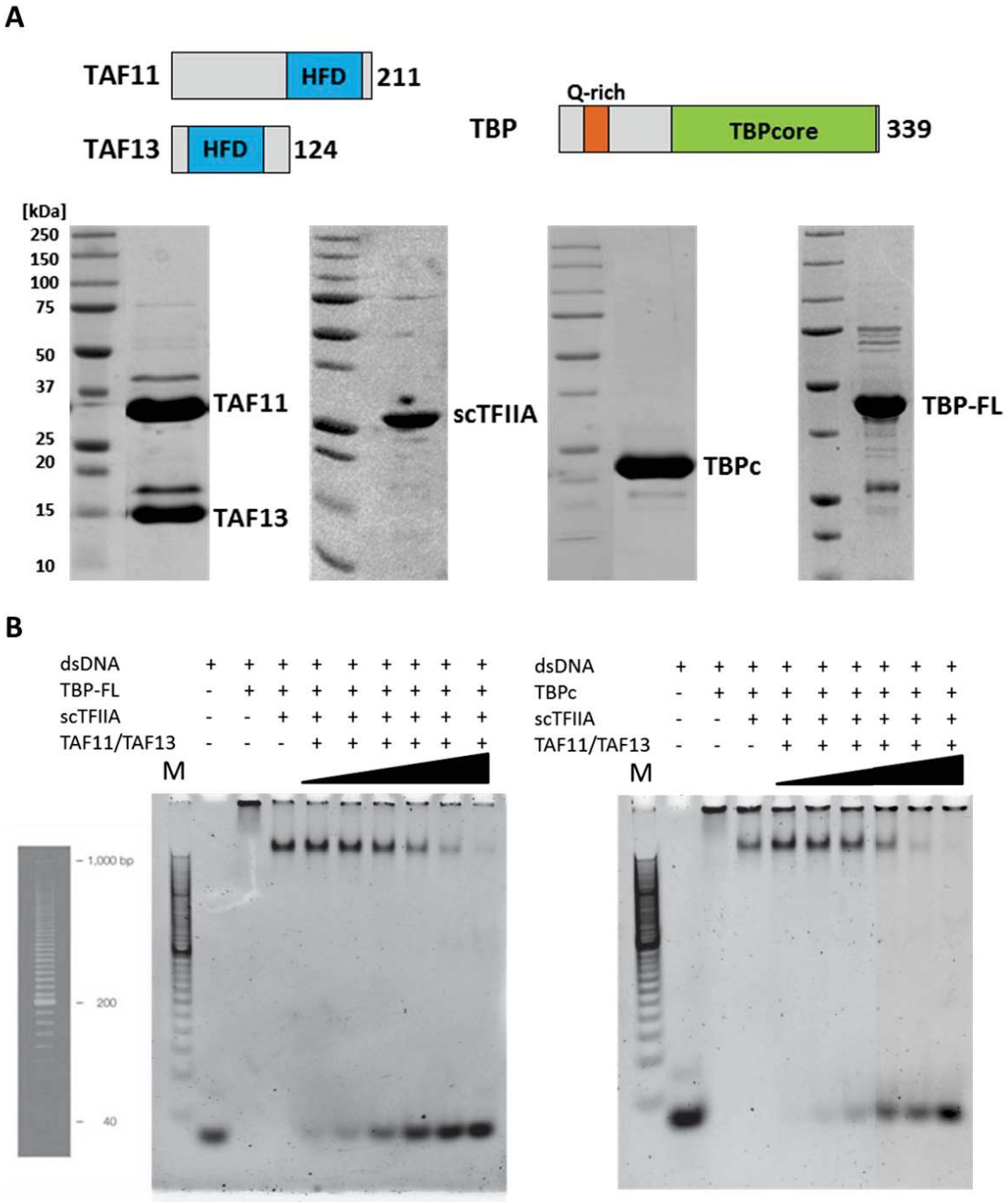


Figure 22: Displacement of DNA from TFIIA/TBP/DNA complex.

(A) Domain representation of TAF11, TAF13 and TBP is shown on top. On bottom SDS-PAGE analysis of the final purified proteins used for EMSA are shown. There were some minor impurities in TAF11/TAF13 and TBP-FL. **(B)** EMSA for verification of hypothesis for pentameric complex formation.

EMSA using TBP-FL is shown on right side and EMSA using TBPc is shown on left side. Shift in electrophoretic mobility indicates the formation TFIIA/TBP/DNA complex (the TBP/DNA complex, lane 3, does not produce a clean shift in the absence of TFIIA). TAF11/TAF13 addition to a preformed TFIIA/TBP/DNA complex results in dissociation of TFIIA/TBP/DNA complex and releases free DNA (Concentration DNA=2μM; TBPc= 2μM; TBP-FL=4μM; scTFIIA= 6μM; TAF11/TAF13= 2-64μM). **M**: Marker

3.2.2. Stoichiometry of the TAF11/TAF13/TBP complex by SEC, Native-MS and AUC analysis

Our results indicated that the pentameric complex was not formed. Moreover, our EMSA results show that free DNA is regenerated when TAF11/TAF13 is added to a preformed TFIIA/TBP/DNA complex. This could have several reasons. For example, it could be that TAF11/TAF13 bind to TFIIA thereby sequestering it from the TFIIA/TBP/DNA complex, which then may be destabilized, dissociating into TBP and releasing free dsDNA. Alternatively, TAF11/TAF13 may bind to TBP and interfere with DNA formation. Then, the TFIIA/TBP/DNA complex would be disrupted and TFIIA, which binds TBP strongly only in the presence of DNA, would also dissociate.

We sought to analyze these different scenarios by biochemical analysis of the interactions between TFIIA, TBP and TAF11/TAF13. SEC analysis of different combinations of the components revealed that a stable complex containing TAF11/TAF13 and TBP could be formed, which would favor the scenario of dissociation of DNA from TFIIA/TBP/DNA by TAF11/TAF13 binding to DNA. SEC analysis of TAF11/TAF13 and TBP on Superdex S200 10/300 column along with SDS-PAGE analysis of elution fractions is shown in Figure 23A. Purified TAF11/TAF13 and TBP were mixed with a molar excess of TBP to form TAF11/TAF13/TBP complex (black). TAF11/TAF13 (red) and TBP (blue) were used as controls. Not that a shift in the elution fractions is not evident from the SEC chromatograms, but SDS-PAGE analysis clearly shows that TBP is co-eluting in the TAF11/TAF13/TBP peak. In fact, TAF11/TAF13/TBP elutes almost at the same position as

previously TAF11/TAF13 eluted. When compared to molecular weight standards (globular proteins) used to calibrate the SEC column, it is apparent that both TAF11/TAF13 and TAF11/TAF13/TBP migrate at larger MWs than expected from adding up the individual molecular weights of the components. This is probably related to TAF11/TAF13 containing long unstructured tails which give rise to enlarged hydrodynamic radii as compared to the globular protein standards (see Appendix and Supplements, Figure 35). The main peak fraction containing the TAF11/TAF13/TBP complex was chromatographed again (shown as green dotted line in Figure 23A) to confirm stability. Analogous experiments, in contrast, evidenced that no complex was formed between TAF11/TAF13 and TFIIA (data not shown).

The TAF11/TAF13/TBP complex was further analyzed by native mass spectrometry (native MS) and analytical ultracentrifugation (AUC). Native-MS analysis was carried out by the lab of Prof. Dame Carol Robinson (Oxford University). The analysis clearly showed the presence of major peaks at ~63 kDa and ~40 kDa. The 63 kDa peak corresponds to the TAF11/TAF13/TBP complex comprising one copy of each subunit, while the 40 kDa peak in the native MS spectra corresponds to a TAF11/TAF13 complex (Figure 23B). AUC analysis confirmed this result, showing a major peak (~75% at 2.1 S) that is consistent with a 1:1:1 heterotrimer containing one subunit each of TAF11, TAF13 and TBP based on the sedimentation value S, the hydrodynamic radii from gel filtration and the method used to prepare the complex. Also in AUC, a smaller second peak (~23% at 3.0 S) was found corresponding to unbound to TAF11/TAF13 complex (Figure 23C). Together, these studies show that TAF11/TAF13 forms a stable complex with TBP with a 1:1:1 stoichiometry containing one copy of each subunit. Of note, TBP, which is a dimer in its DNA unbound form, is present as a monomer in the TAF11/TAF13/TBP complex.

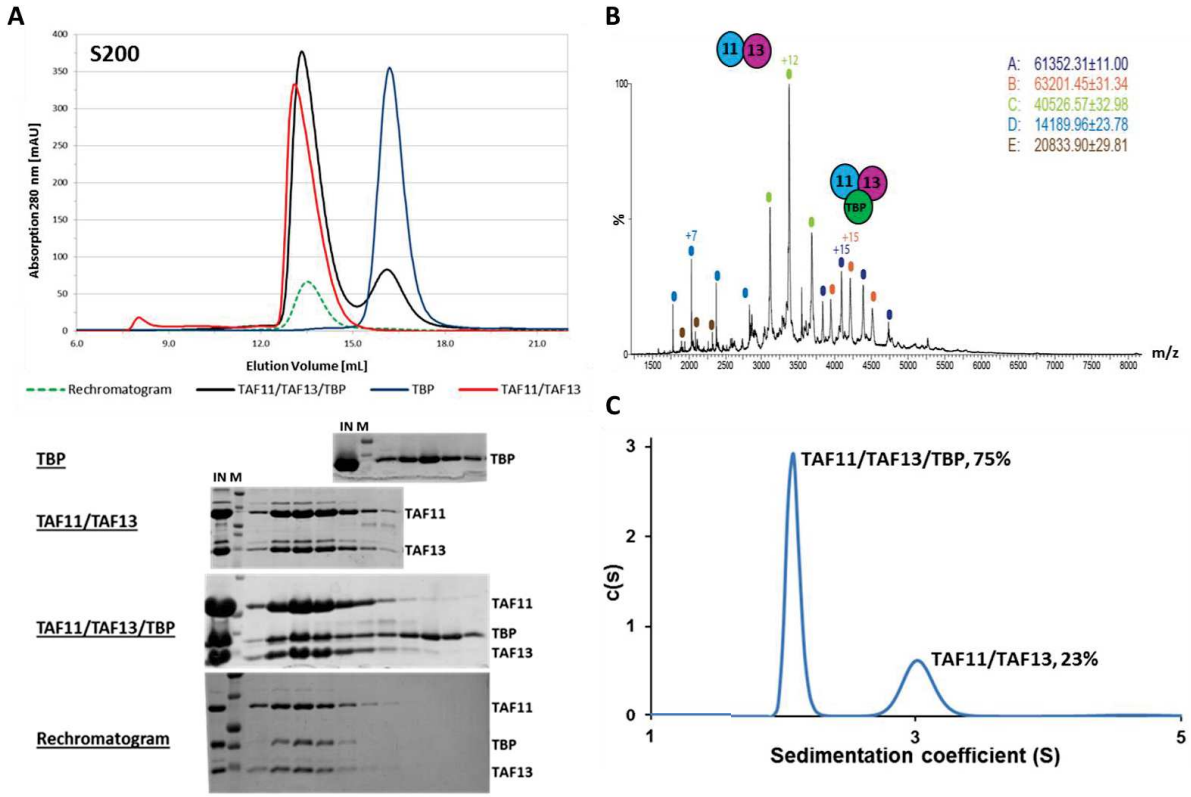


Figure 23: Formation and Characterization of TAF11/TAF13/TBP complex.

(A) Size exclusion chromatograms on Superdex S200 10/300 column and SDS-PAGE analysis indicating the formation of TAF1/TAF13/TBP complex. There is no peak shift but SDS-PAGE shows the shift of TBP into TAF11/TAF13 peak. **IN**: Input sample, **M**: Marker. **(B)** Native MS analysis for TAF11/TAF13/TBP complex indicates for the presence of the peaks at ~60 kDa which corresponds to a 1:1:1 hetero-trimeric complex of TAF11, TAF13 and TBP. Peaks at ~40 kDa and 20 ~kDa corresponds to TAF11/TAF13 and TBP, which might be coming from the dissociation of the complex during experiment (Expected molecular weight for equimolar TAF11/TAF13/TBP complex and TAF11/TAF13 complex are 63017 Da and 40672 Da). **(C)** Analytical ultracentrifugation analysis for TAF11/TAF13/TBP indicates for an elongated 1:1:1 hetero-trimeric complex of TAF11, TAF13 and TBP. Second peak corresponds to TAF11/TAF13 impurity.

3.2.3. TATA box mimicry by TAF11/TAF13?

Formation of a stable complex between TAF11/TAF13 and TBP is in agreement with the literature where TAF11 and TAF13 were found to interact with TBP shown by

immunoprecipitation (May et al., 1996; Mengus et al., 1995). Displacement of DNA by TAF11/TAF13 from a TFIIA/TBP/DNA complex further suggests that the binding surface on TBP for DNA and TAF11/TAF13 might be identical or overlapping. Indeed, previously a TATA box "mimicry" of the N-terminal TBP binding domains of TAF1 (TAF1-TAND) was shown, where TAF1-TAND binds to the DNA binding concave surface of TBP and mimics TATA-box DNA (Liu et al., 1998; Mal et al., 2004) (see also Chapter 1.4.4). Human TAF1-TAND can be divided in two separate domains (TAND1 and TAND2) as shown in Figure 24. We investigated whether these domains compete with TAF11/TAF13 for TBP binding or not, with the aim to find out whether or not TAF11/TAF13 binding to TBP and TAF1-TAND binding to TBP are mutually exclusive. TAF1-TAND1 and TAF1-TAND1+2 domains were produced with an MBP-tag at the N-terminus in baculovirus infected insect cells. Amylose-resin pull-down assay was performed, as described in Methods (Chapter 2.2.6.10), to analyze competition of these domains with TAF11/TAF13 for the DNA binding surface of TBP. As shown in Figure 24, when preformed TAF11/TAF13/TBP complex is added to TAF1-TAND1+2 immobilized on amylose resin via their MBP-tag, it results in displacement of TBP from TAF11/TAF13/TBP complex. TAF11/TAF13 elutes in the flow-though, whereas TBP associates with TAF1-TAND1+2. We also found that, in an analogous experiment. TAF1-TAND1 alone was not sufficient to retain TBP. These results suggests that TAF11/TAF13 and TAF1-TAND binding to TBP is mutually exclusive i.e. that the proteins may share (parts of) the interaction surface on TBP.

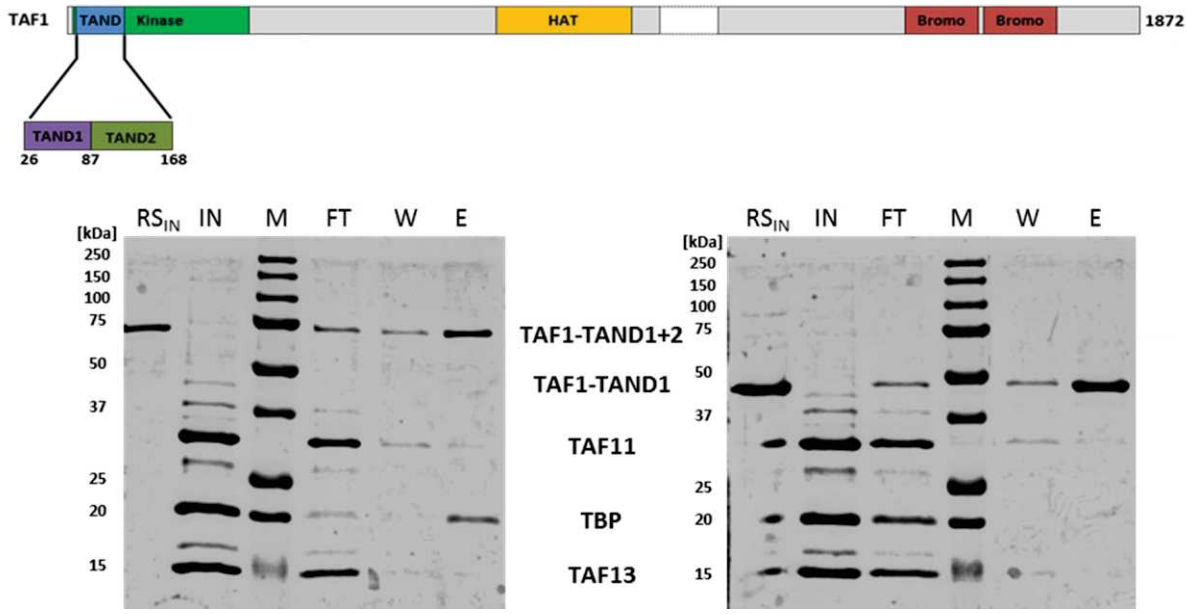


Figure 24: Competition of TAF11/TAF13 with TAF1-TAND for TBP binding.

Domain representation for TAF1-TAND is shown on top. SDS-PAGE analysis for MBP pull down assay using MBP-TAND12 (left) and MBP-TAND1 (right), indicates that TAF1-TAND competes with TAF11/13 for TBP binding and displaces TBP from preformed TAF11/TAF13/TBP complex. Small amounts of TAND domains, were observed in flow through and washes as well upon addition of TAF11/TAF13/TBP complex; indicating leakage from the resin. **RS$_{IN}$:** Resin input (with MBP-TAND12 or MBP-TAND1), **IN**: Input sample (preformed TAF11/TAF13/TBP), **M**: Marker, **FT**: Flow through, **W**: wash, **E**: Elution.

3.2.4. Crystallization trials of TAF11/TAF13/TBP complex

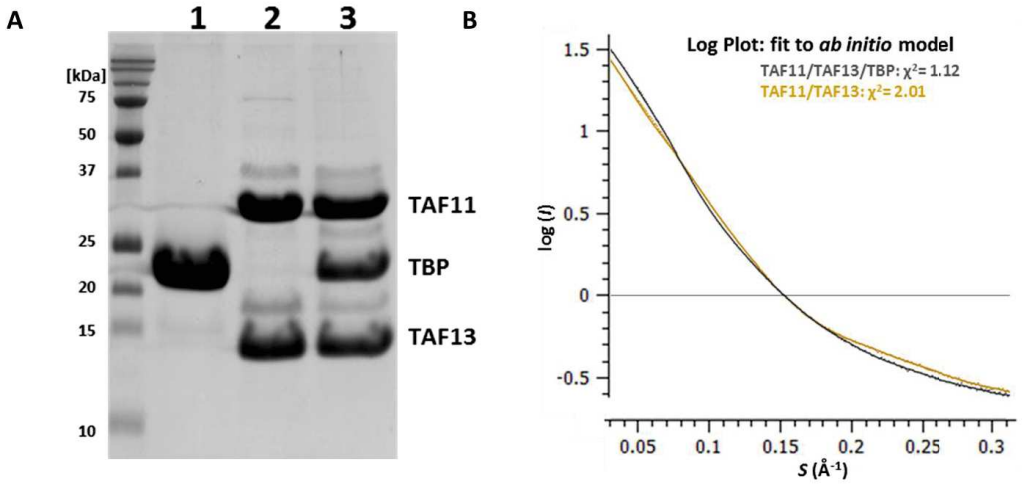
Our results suggest that TAF11/TAF13 may bind to TBP via its DNA binding surface in a mode of TATA-box mimicry and structural characterization of this complex can provide a detailed view of this interaction at the amino acid level, thus providing a basis to probe this interaction by mutational analysis *in vitro* and *in vivo*. In order to do so, TAF11/TAF13/TBP complex containing full-length TAF11 and TAF13 and the conserved core of TBP was purified to homogeneity as described in Chapter 3.2.2 for crystallization studies. Only the central peak fractions were used for crystallization screening using nanoliter high-throughput

crystallization facility at the HTX laboratory (EMBL Grenoble). Close to 900 crystallization conditions at two different protein concentrations were screened covering a broad range of crystallization space, however no crystals were obtained to date. Long unstructured tails are present in TAF11 and TAF13, which may be preventing crystal formation. Thus, to enable crystallization, a series of truncation constructs of both TAF11 and TAF13 was produced to remove (parts of) the unstructured tails, based on secondary structure predictions. These truncated complexes were purified to homogeneity using the protocols established for the full-length specimen and checked for interaction with TBP, similarly as was done for full length TAF11/TAF13. All complexes were subjected to crystallization screening. These truncation construct are shown in Figure 36 in Appendix and Supplements. Crystallization trials were ongoing at the time this thesis was written.

3.2.5. SAXS analysis of TAF11/TAF13/TBP complex

The structure of TBP core has been determined at atomic resolution (Kim et al., 1993a). Likewise, the crystal structure of a portion of the TAF11/TAF13 histone fold dimer has been elucidated (Birck et al., 1998), Therefore, we reasoned that global structural information could be gathered by using these atomic coordinates in conjunction with a molecular envelope determined by SAXS. Thus, we subjected the TAF11/TAF13/TBP to analysis by SAXS. Main peak fractions of purified TAF11/TAF13/TBP complex and TAF11/TAF13 along with TBP as control were used to collect static SAXS data at the ESRF BioSAXS beamline BM29 equipped with a Pilatus 1M detector (with Dr. Adam Round). Scattering curves were collected for different concentrations of samples with 10 frames for each concentration (see Appendix and Supplements, Table 14). Data was further analyzed as described in Methods (Chapter 2.2.6.12). Scatterings from different concentrations (constant Rg values) were merged to decrease noise. Data analysis suggested for no aggregation as shown by Guinier plot, while Kratky plot indicated for the presence of flexibility in both TAF11/TAF13 and TAF11/TAF13/TBP (see Appendix and Supplements, Figure 37).

Data quality was good as shown by different classical parameters used for validation (see Appendix and Supplements, Table 14). Rg calculated from Guinier analysis were 40.9 ± 0.6 Å and 40.3± 3.6 Å for TAF11/TAF13/TBP and TAF11/TAF13 complexes respectively, while Dmax were 160 Å and 140 Å respectively. *Ab initio* models were calculated with good χ^2 values of 1.12 and 2.01, respectively for TAF11/TAF13/TBP complex and TAF11/TAF13 complex and with a nice fit to the data as described in Methods (Chapter 2.2.6.12) (Figure 25B). The *Ab initio* model of TAF11/13 was nicely overlapped with that of TAF11/TAF13/TBP using supcomb in ATSAS software package, leaving some extra density present in TAF11/TAF13/TBP, as shown in the Figure 25C. In this extra density, TBP was fitted manually in PyMOL and only one copy of TBP (shown as green cartoon) can be fitted. Available structure of TAF11/TAF13 (shown as blue and magenta cartoon) was also placed in these *ab initio* models using supcomb and PyMOL. This analysis provides a low resolution structural data which will be further used with more structural data to create a model for this interaction.



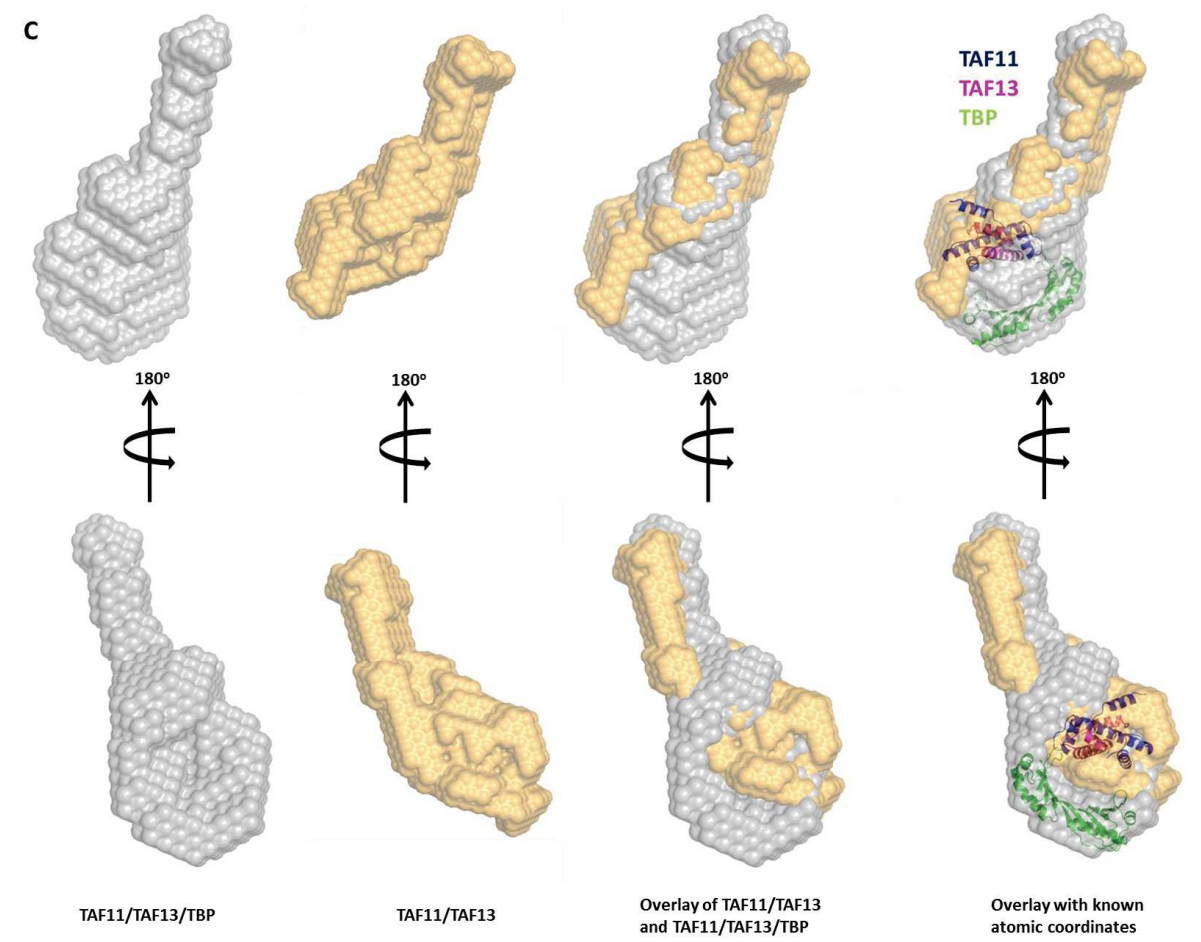


Figure 25: SAXS analysis of TAF11/TAF13/TBP complex.

(A) Purified TBP (1), TAF11/TAF13 complex (2) and TAF11/TAF13/TBP complex (3) used for SAXS analysis. **(B)** Log plot for fit of experimental scattering (shown as dots) to the calculated scattering from ab initio models (shown as solid lines). TAF11/TAF13 complex is shown in light orange color and TAF11/TAF13/TBP complex is shown in light grey color, as surface representation. Error bars are shown as vertical grey lines. Corresponding $\chi 2$ values for different models are also shown. **(C)** *Ab initio* model for TAF11/TAF13/TBP (grey surface), for TAF11/TAF13 (light orange surface), overlay of both *ab initio* models and placement of crystal structures of TBP (green cartoon, PDB ID **1CDW**) and TAF11/TAF13 (blue and magenta cartoon, PDB ID **1BH8**) in these *ab initio* models are shown from two different angles. This analysis shows that TAF11/TAF13 model nicely fits with TAF11/TAF13/TBP model, leaving a space for only one copy of TBP.

3.2.6. CLMS analysis of TAF11/TAF13/TBP complex

The SAXS envelope is highly suggestive of TAF11/TAF13 interacting with TBP via the DNA binding interface, in good agreement with all our previous observations. However, the orientation of TBP relative to TAF11/TAF13 cannot be unambiguously deduced from the SAXS analysis alone, due to low resolution. To obtain more information if possible directly from the interaction interface, we set out to analyze the TAF11/TAF13/TBP complex by chemical cross-linking mass spectrometry (CLMS), with the objective to obtain distance restrains to define the TAF11/TAF13 interaction with TBP. Purified TAF11/TAF13/TBP complex and TAF11/TAF13 complex (as control) were chemically cross-linked using BS3 followed by a SEC purification using Superdex S200 10/300 (Figure 26A) and analyzed by mass spectrometry as described in Methods (Chapter 2.2.6.13) in collaboration with Juri Rappsilber (University of Edinburgh). The band for cross-linked TAF11/TAF13/TBP was excised from a native PAGE gel for mass spectrometry analysis, to forestall contamination of cross-linked TAF11/TAF13 or remaining uncross-linked proteins. A number of crosslinks were obtained for TAF11/TAF13/TBP complex (Figure 26B) (also see Table 15 in Appendix and Supplements). This preliminary cross-linking data suggests that indeed, TAF11/TAF13 mainly crosslinks towards the DNA binding surface of TBP, supporting our previous observations and conclusions. Crosslinks for N-terminal region of TAF11 were markedly missing; reason being that it is devoid of amino acid lysine, which is required for cross-linking by BS3. Distance restrains obtained from this cross-linking data along with SAXS data and existing crystal structures were used to obtain a preliminary atomic model of the TAF11/TAF13/TBP complex by using algorithms developed by Andrej Sali (UCSF, USA), and integrated multi-parameter modeling in collaboration with Dr. Aleksandra Watson and Prof. Ernest Laue (University of Cambridge) (Figure 26C). The model comprises of a compact core formed by HFD of TAF11/TAF13 (blue and magenta) and TBP (green) followed by long unstructured tail of TAF11. Different lysine residues involved in crosslinking are shown in Figure 26D. This preliminary model satisfies more than 90% of

cross links found in CLMS experiments as well as the SAXS data. Crosslinks between TAF11 and TAF13 also satisfies the crystal structure of TAF11/TAF13 histone fold domains (Birck et al., 1998). Further analysis and refinement is ongoing at the time this thesis is written.

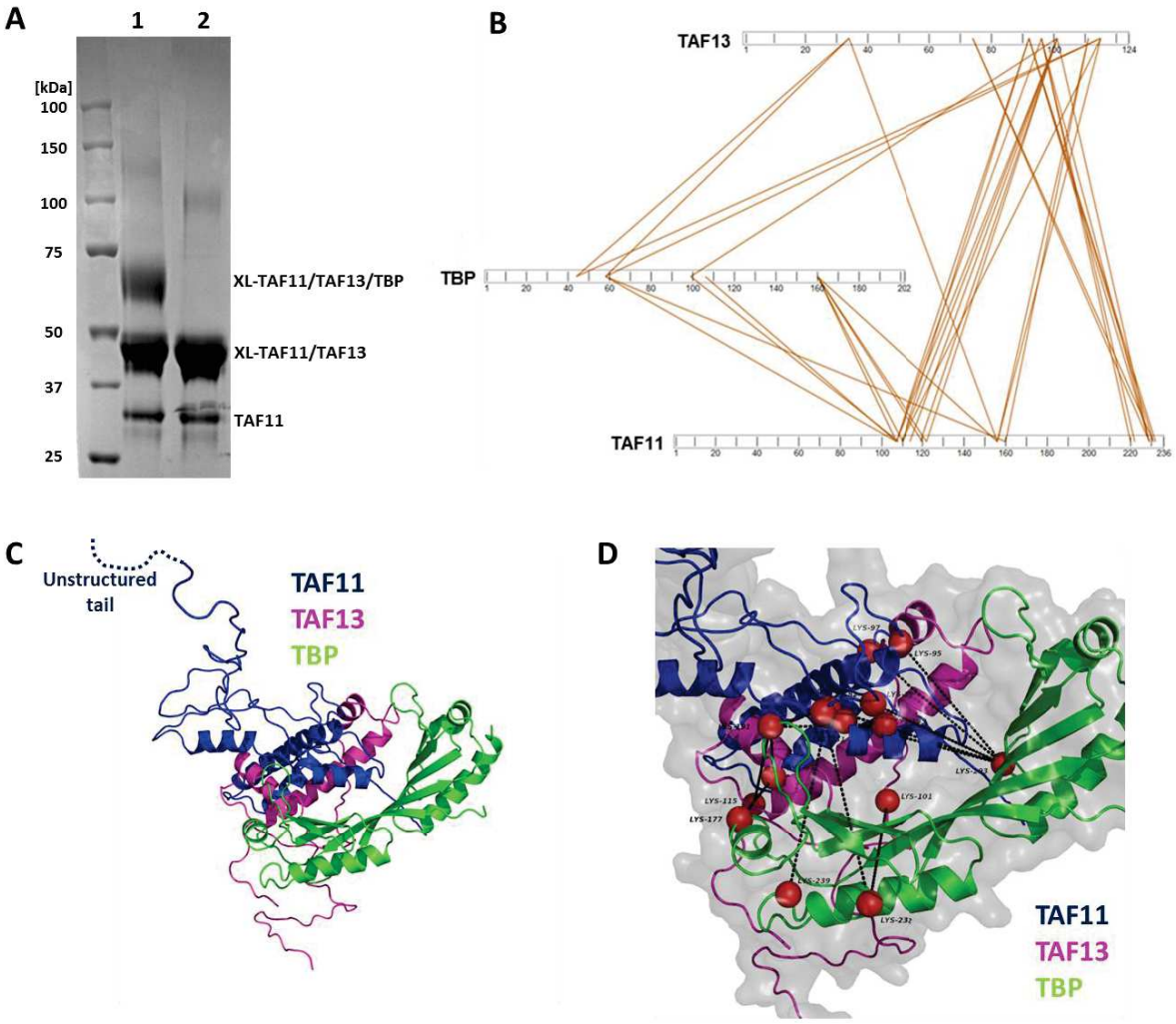


Figure 26: CLMS analysis of TAF11/TAF13/TBP complex.

(A) SDS-PAGE analysis of cross-linked (with BS3) and purified TAF11/TAF13/TBP complex (1) and TAF11/TAF13/TBP complex (2) is shown. Top band corresponds to cross-linked TAF11/TAF13/TBP complex, while middle band corresponds to TAF11/TAF13 complex. The band on the bottom corresponds to uncross-linked TAF11. **(B)** Crosslinks between TAF11-TAF13, TAF11-TBP and TAF13-TBP are shown with 5% FDR cutoff data. No cross links are shown for N-terminal of TAF11, as it is lacking amino

acid Lys. **(C)** Preliminary quasi atomic model of TAF11/TAF13/TBP complex derived from integrated multi-parameter modeling. TAF11 (blue), TAF13 (magenta) and TBP (green) makes a compact core which is flanked by unstructured tails of TAF11 and TAF13. **(D)** Different lysine pairs involved in crosslinking are shown (red spheres) connected with dotted black lines.

3.3. Structural characterization of 9TAF complex

9TAF is a 1.3 MDa intermediate subcomplex in TFIID assembly, formed by addition of TAF3/TAF10 to 8TAF complex, as described earlier (see Chapter 1.4.4 & 1.5.) Structural studies of this 9TAF complex might provide more detailed information about TFIID assembly and also it can help us to locate TAF3 within this complex which also might reveal some function of this TAF in TFIID assembly.

3.3.1. Preparation and negative stain RCT data collection of 9TAF

Our current reconstitution protocol of fully recombinant human holo-TFIID with a full complement of TAFs and TBP relies on a reconstitution step in which a highly purified so-called 9TAF complex comprising TAF2, TAF3, TAF4, TAF5, TAF6, TAF8, TAF9, TAF10 and TAF12 is mixed with a so-called MBPTAF1-Module comprising TAF1, TAF7, TAF11, TAF13 and TBP to yield active holo-TFIID complex. The previous two parts of the Results section of this thesis (Chapter 1.1 and 1.2) addressed architectural features of TAF1 and TAF7 on the one hand, and a novel TAF11/TAF13/TBP complex on the other, which are related to the MBPTAF1-Module. The current chapter describes the analysis of the 9TAF constituent that together with the MBPTAF1-Module gives rise to holo-TFIID *in vitro*.

The 9TAF complex was prepared to homogeneity by following established protocols in the Berger laboratory as described in PhD thesis of Yan Nie, 2012 (Nie, 2012). The 9TAF complex was stabilized by mild cross-linking using glutaraldehyde, applying the GraFix method (Stark, 2010). To improve purity of the complex, GraFix was followed by a size exclusion chromatography step. Highly purified 9TAF complex was then used to prepare

negatively stained grids using 2% uranyl acetate as stain. These grids were used to collect EM data for a Random Conical Tilt (RCT) reconstruction as described in the Methods (Chapter 2.2.7.2) (Figure 27).

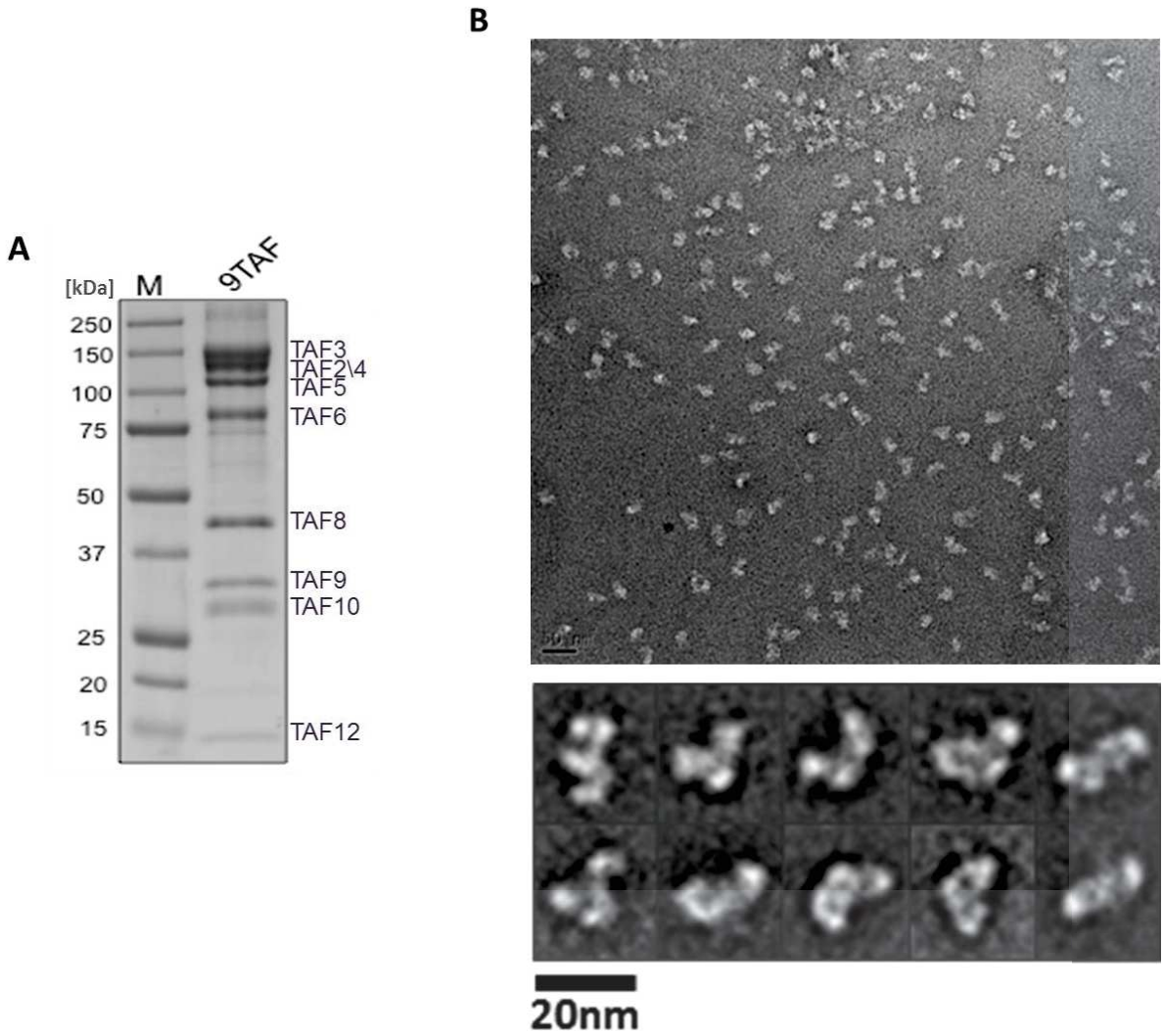


Figure 27: Negative stain EM and 2D MSA of 9TAF complex.

(A) Coomassie Brilliant Blue stained SDS-PAGE gel section of highly purified 9TAF complex. Subunits are indicated. M stands for marker. **(B)** Negative stain micrograph of GraFix treated and SEC purified 9TAF complex is shown (top) along with representative 2D class averages (bottom).

3.3.2. Negative stain EM analysis of 9TAF

All micrographs were pre-processed using IMOD and Bsoft (Heymann and Belnap, 2007) to remove bad image points (from X-ray) and lines (from camera imperfection), and then binned by a factor of 2 by Bsoft. These preprocessed micrographs were then evaluated by contrast transfer function (CTF) estimation using XMIPP software packages (Sorzano et al., 2004). A total of 10313 particle pairs were manually picked from 203 micrograph pairs using TiltPicker (Voss et al., 2009). These particle pairs were extracted from micrographs using SPIDER (Shaikh et al., 2008) and XMIPP and then preprocessed (particle normalization, ramping background correction, and band-pass filtering) using XMIPP. The untilted views of these particle pairs were analyzed by 2D MSA of IMAGIC (van Heel et al., 1996) to generate total 250, 2D classes. Highly populated classes are shown in Figure 27B. All classes represent different orientations of the complex exhibiting distinct features. These 2D classes were used for 3D reconstruction of 9TAF by RCT method. 9TAF particle were classified by ML2D of XMIPP using 2D classes from IMAGIC as reference and RCT 3D models were reconstructed for each class using ML3D of XMIPP. These classes were averaged to 25 3D volumes using MLtomo of XMIPP. Out of these, 4 models were selected which comprised most of the 2D classes. Additional 11958 particles were picked manually from 168 untilted micrographs using e2boxer of EMAN2 (Tang et al., 2007), and added to 10313 untilted particles from tilt pairs, summing up to a total of 22271 particles.

These particles were used to further refine the four selected models of the 9TAF complex using MLrefine3D of XMIPP. Refined RCT models are shown in Figure 28. All models were verified by presence of reprojections from these models (generated using SPIDER) in 2D classes as shown in Figure 28. The models suggest that 9TAF is dynamic and that several distinct conformations are present in the sample used for RCT.

Intriguingly, one of these models (Fig 27, Right) already adopts a structure reminiscent of the clamp shape exhibited by the low-resolution reconstructions of endogenous TFIID (Andel et al., 1999; Brand et al., 1999; Elmlund et al., 2009; Grob et al.,

2006). These models, which I calculated, provide initial reference models for high-resolution 3D structure determination of 9TAF complex by cryo-EM. The four RCT models will also be useful for sorting particles to discern conformational heterogeneity, if present, in future cryo-EM data.

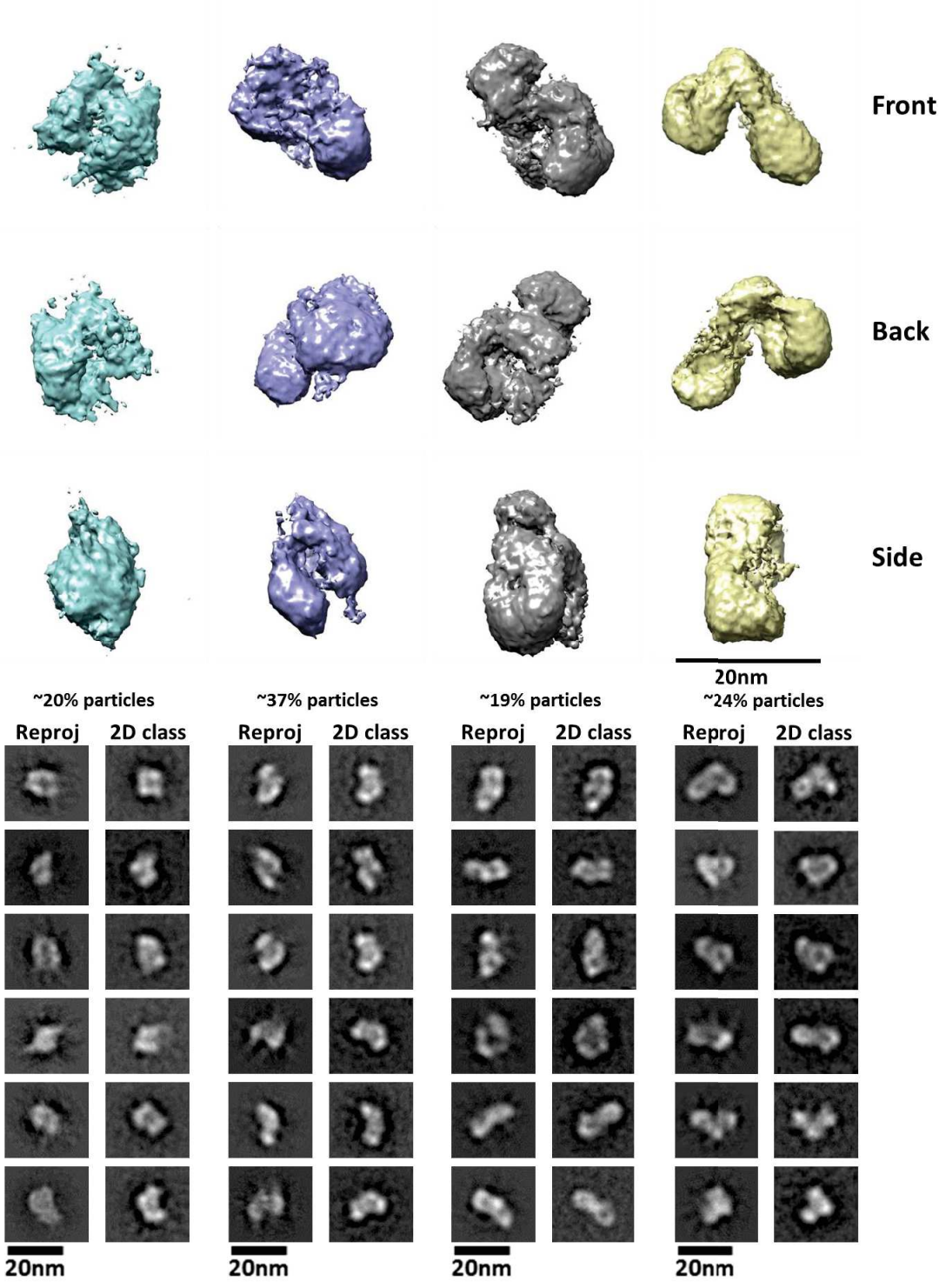


Figure 28: RCT models of highly purified 9TAF complex.

Four refined RCT models of 9TAF complex are shown from front, back and side views. Amount of particles present in each class is also shown. On the bottom comparison of re-projections with 2D classes is shown as validation.

4. DISCUSSION AND FUTURE PERSPECTIVE

SUMMARY IN ENGLISH

In this Chapter of my thesis, I will discuss the results obtained and their potential and significance. In the present thesis, three particular project lines were pursued. In Part I, by combining a range of structural biology and biophysical methods, including X-ray crystallography, NMR and SAXS, the structure of the complete human TAF1/TAF7 interaction was analyzed. These studies revealed structural features which are evolutionary conserved and particular aspects which are unique to the human TFIID components. Of note, the structures we and others obtained question earlier reports which postulated a histone acetylase domain within TAF1. Moreover, our findings may have implications of the assembly of tissue specific forms of TFIID in humans, including for instance a TFIID variant containing specific isoforms of TAF1 and TAF7 which has key functions in spermatogenesis.

Part II of this thesis concerned a novel TAF11/TAF13/TBP complex which we fortuitously discovered when attempting to reproduce published data on a putative pentameric TAF11/TAF13/TFIIA/TBP/DNA complex. This novel complex was characterized by a combination of structural and biochemical means, and a first quasi-atomic model of the complex was obtained by applying hybrid methods. This complex may represent a new case of TATA-box mimicry, with a protein specimen occupying the DNA binding site of TBP, thus preventing TBP-DNA interactions. TATA-box mimicry was previously observed in a complex between the N-terminal domains of TAF1 and TBP, and in a complex between the transcription factor Mot1 and TBP.

Part III finally discussed first negative-stain EM reconstructions of 9TAF complex using random conical tilt (RCT) approach. 9TAF contains 9 out of 13 TAFs in TFIID and represents one of two building blocks we use in the Berger laboratory to reconstitute fully recombinant and active holo-TFIID. Notably, 9TAF contains all TFIID subunits that are thought to be present in two copies in TFIID. This 9TAF complex may represent an assembly

intermediate of TFIID in the cell, which may acquire the remaining TAFs and TBP on a putative assembly pathway when holo-TFIID is formed in the cell nucleus.

RÉSUMÉ EN FRANÇAIS

Dans ce chapitre, je vais discuter les résultats obtenus de leur potentiel et de leur importance. Cette thèse portée sur trois projets. Dans la partie I, des méthodes biophysiques et des méthodes de biologie structurale (cristallographie par rayons X, RMN et SAXS) ont étaient combinées pour analyser l'intégralité de la structure du domaine d'interaction du complexe TAF1/TAF7 humain. Cette étude a révélé des caractéristiques structurelles conservées au cours de l'évolution et des aspects particuliers qui sont uniques aux composants de TFIID humain. Il est important de noter, que la structure que nous présentons ainsi qu'une autre équipe questionne des rapports antérieurs qui postulait la présence d'un domaine histone acétylase au sein de TAF1. En outre, nos résultats pourraient avoir des implications sur les différents assemblages possible de TFIID spécifique de certain tissus chez l'homme, notamment pour le cas d'un variant de TFIID contenant des isoformes spécifiques de TAF1 et TAF7 qui a des fonctions de clé dans la spermatogénèse.

Dans la seconde partie de cette thèse nous traitons d'un nouveau complexe TAF11/TAF13/TBP que nous avons découvert en tentant de reproduire des données publiées sur un complexe pentamérique putatif TAF11/TAF13/TFIIA/TBP/ADN. Ce nouveau complexe a été caractérisé par la combinaison d'analyses structurelles et biochimiques, et un premier modèle quasi-atomique du complexe a été obtenu par l'application de méthodes hybrides. Ce complexe peut représenter un nouveau cas de mimétisme de la boîte TATA, avec une protéine occupant le site de liaison à l'ADN de TBP empêchant ainsi les interactions TBP/ADN. Le mimétisme de la boîte TATA a été précédemment observé dans un complexe entre les domaines N-terminaux de TAF1 et TBP, et un complexe entre le facteur de transcription MOT1 et TBP.

La troisième partie de cette thèse, discute en premier lieu de la reconstruction du complexe 9TAF par coloration négative en électromicroscopie utilisant une approche d'inclinaison conique aléatoire (RCT). 9TAF contient neuf des treize TAFs de TFIID et représente l'un des deux blocs de construction que nous utilisons dans le laboratoire du Pr Berger pour reconstituer l'intégralité de l'holo-TFIID actif. Notamment, 9TAF contient toutes les sous-unités de TFIID qui sont présentes en deux exemplaires dans TFIID. Ce complexe 9TAF peut représenter un ensemble intermédiaire de TFIID dans la cellule, qui ensuite pourrait acquérir les TAFs restants et TBP au cours d'une voie de montage putative pour la formation de l'holo-TFIID dans le noyau de la cellule.

4.1. TAF1/TAF7 interactions

A set of studies were carried out with the objective to decipher the structure and dynamics of human TAF1/TAF7 complex formation. I co-expressed and purified various versions of TAF1 and TAF7 which contained the entire proposed interaction region or parts of this region, respectively. I determined the 1.75Å crystal structure of a partial human TAF1/TAF7 complex. In addition, I collaborated with Martin Blackledge and Malene Ringkjøbing Jensen to determine the NMR structure of a small high-affinity TAF1/TAF7 interaction motif that was not present in my crystal structure, and furthermore was connected to the crystallized domains by a long, non-conserved and unstructured linker. The X-ray coordinates and the NMR solution structure were both modelled into a molecular envelope derived from SAXS experiments to yield a molecular description of the complete human TAF1/TAF7 interaction. Consistent with structural results from other laboratories that became available at the same time, our data show that the TAF1/TAF7 interaction is highly intricate and involves a shared β-barrel motif with TAF1 and TAF7 both donating β-strands that are intertwined. The crystal structure of a yeast TAF1/TAF7 complex was also determined recently (Bhattacharya et al., 2014). Comparison of human TAF1/TAF7 and yeast TAF1/TAF7 shows similarities and differences. Yeast TAF1/TAF7 appears to be a single compact interaction unit comprising the β-barrel and two arm domains. In contrast, the human TAF1/TAF7 interaction appears to be bipartite and can be divided in two distinct N- and C-terminal interaction regions. The N-terminal interaction region contains the intertwined shared β-barrel and one arm domain and is contained in our crystal structure. The C-terminal region is formed by a small high-affinity interaction motif (Kd=100-200 nM), and our NMR structure shows that it corresponds to the second arm domain in yeast TAF1/TAF7. In contrast to yeast, this domain in human is however attached to the β-barrel by an extended non-conserved linker domain encompassing ~50 amino acids. Based on our SAXS data, this small high affinity interaction domain appears to be more flexibly attached to the β-barrel unit, which is consistent with the presence of the flexible linker. This flexibility may also explain why we could not crystallize

the entire TAF1/TAF7 interaction region. Notwithstanding, our hybrid approach by combining SAXS, NMR and X-ray data allowed us to derive a composite hybrid model of the complete human TAF1/TAF7 interaction.

The β-barrel structure is conserved in yeast and human TAF1/TAF7, and is closely reminiscent of the β-barrel found in human TFIIF, which is a heterodimer formed by the subunits RAP74 and RAP30. In TFIIF, RAP74 and RAP30 both also donate β-strands to form a similar β-barrel (Gaiser et al., 2000). Moreover, a recent study showed that a transcription factor complex, TFIIIC, which is involved in regulating transcription by Pol III, also contains a similar β-barrel fold (Taylor et al., 2013). These findings indicate that this motif may be prevalent in transcription factor complexes, and could suggest a functional and evolutionary conservation between TFIIF, TFIIIC and TFIID, although it will require further study to understand what exactly the implications of such a conservation would be.

Intriguingly, the TAF1 segment present in the TAF1/TAF7 complex structure contains a winged helix domain. Winged helix domains are known to mediate DNA binding. The TAF1 winged helix domain is similar to the winged helix domain of RAP30 in human TFIIF which can bind DNA (Groft et al., 1998). TAF1 has previously been shown to have a role in core promoter binding (Verrijzer et al., 1995) which may be mediated by this winged helix domain. In fact, recent biochemical experiments indicate that the TAF1 winged helix domain can indeed bind DNA (Wang et al., 2014).

Interestingly, the TAF1 domain used in this study contains a region that was proposed to have HAT activity previously. However, the corresponding TAF1 architecture in the TAF1/TAF7 complex does not at all resemble any known HAT domain. It was suggested previously that TAF7 may inhibit the acetyltransferase activity upon binding to TAF1. Thus, it cannot be ruled out that binding of TAF7 to TAF1 wholly reconfigures a folded HAT enzymatic domain within TAF1 into the geometry observed in the TAF1/TAF7 complex, thereby inhibiting it, but the reorganization would be massive. Recently, several previously proposed activities of TAF1 came under scrutiny as doubts surfaced about the validity of

these early studies. The structural results presented here substantiate reservations that TAF1 actually contains a HAT or AT domain. The possibility exists that this enzymatic activity may need to be re-attributed to contaminations in these early and crude TAF preparations.

The presence of intertwined β-barrel in N-terminal interaction region suggests that co-folding of these two proteins, TAF1 and TAF7, is required to form this intricate structure. In our laboratory, the current protocol for holo-TFIID production we produce holo-TFIID is by reacting highly purified 9TAF complex with a so-called "MBPTAF-Module" which comprises the remaining TAFs and TBP. This protocol results in holo-TFIID which is complete and active in transcription assays and in EM has the characteristic clamp shape known from EM studies of TFIID purified from endogenous source (See Figure 29). We occasionally observed that TAF7, which is added as a purified entity to immobilized TAF1 to nucleate the "MBP-TAF module", was present in varying amounts in holo-TFIID from preparation to preparation (data not shown). The TAF1/TAF7 structure implies that these two proteins need to be co-expressed, and we are incorporating this approach in a modified holo-TFIID purification protocol to improve the quality of recombinant holo-TFIID we are investigating. In this modified protocol co-expressed TAF1/TAF7 will be used to nucleate MBPTAF-Module; which will be then further mixed with 9TAF to form holo-TFIID.

Evidences suggest that TAF7 may be released from PIC after transcription initiation, when Pol II transits to transcript elongation (Gegonne et al., 2013; Gegonne et al., 2006). Phosphorylation of TAF7 by a kinase activity contained within TAF1 was proposed to disrupt the TAF1/TAF7 interaction which is thought to result in TAF7 release from PIC. Specific phosphorylation site(s) in TAF7 were found in its C-terminal part (Gegonne et al., 2001; Gegonne et al., 2006; Kloet et al., 2012). Interestingly, these are not present in the proposed interaction region. Based on the crystal coordinates, it is unclear how the tight interactions in the TAF1/TAF7 complex would be disrupted by phosphorylation of TAF7, given that the proposed phosphorylation of TAF7 is not within the β-barrel region which would need to be unfolded. In the experiments that were previously carried out to show

dissociation of TAF7 from TAF1 by TAF7 phosphorylation, both proteins were separately expressed and the complex then reconstituted by mixing. From our results, it would appear that in these experiments, probably only the small high-affinity interaction motif was formed, while the intertwined TAF1/TAF7 β-barrel was not. The small high-affinity interaction domain, as evidenced by our NMR structure, can be reconstituted by mixing, and is presumably strong enough to keep full-length TAF1 and TAF7 together even in the absence of a properly folded shared β-barrel. We assume that it is in fact this small high-affinity interaction motif which was disrupted by phosphorylation of TAF7, leading to dissociation of the *in vitro* assembled TAF1/TAF7 complex. Clearly, further experiments are required using co-expressed TAF1/TAF7 to clarify the mechanisms of transcription activation by TAF7 release from TFIID.

The human TAF1/TAF7 complex resembles the recently determined yeast TAF1/TAF7 structure notably in the β-barrel region, providing evidence for a functional conservation from yeast to human, and identifies highly complicated and conserved interaction between both proteins. The major difference between human and yeast TAF1/TAF7 is the presence of a ~50AA long partially unstructured region in TAF1 between the β-barrel domain and the small high-affinity interaction motif. This leads to increased flexibility in human TAF1/TAF7 as compared to the yeast complex, which is further supported by our SAXS data. The SAXS data clearly shows that crystal and NMR structures can be fitted nicely within *ab initio* models and confirms the flexibility. Our peptide array analysis, finally; corroborated specific TAF7 interactions with TAF1 at single residue acid level, which could be used for mutational studies *in vivo* in the future.

In summary, the results presented, convey a requirement of both proteins, TAF1 and TAF7, to maintain each other's structural integrity by forming the TAF1/TAF7 complex, presumably by co-folding. This is of particular interest in the context of paralogues of TAF1 and TAF7 which define distinct TFIID complexes in a tissue specific manner. Subunit allocation (allocation of subunits in different cellular compartments; which then comes

together to form the complete complex) is emerging as a key concept of transcription regulation, determining differentiation and development programmings, but the underlying mechanisms are currently not understood. For example, in spermatogenesis, a distinct TFIID complex is found, which contains the paralogues TAF1L and TAF7L but not TAF1 or TAF7 (Pointud et al., 2003; Wang and Page, 2002). How does the cell ensure that only the paralogues are allocated to this tissue specific TFIID variant? What are the atomic determinants that exclude the paring of TAF1 with TAF7L, or TAF1L with TAF7, respectively? It will be interesting to investigate the paralogue TAF1L/TAF7L complex, to understand the tissue specific presence of these and other paralogues in distinct TFIID complexes, and the functional consequences.

4.2. 'TATA-box mimicry' by TAF11/TAF13?

It was shown previously that a TBP/DNA complex is stabilized by TFIIA, resulting in a stable ternary TFIIA/TBP/DNA complex. Based on biochemical and genetic data, it was proposed that this TFIIA/TBP/DNA complex is further stabilized by TAF11 binding. TAF11 was later found to form a highly stable complex with TAF13, in which TAF11 and TAF13 interact tightly via their histone fold domains. Together, this data suggested the presence of a stable pentameric TAF11/TAF13/TFIIA/TBP/DNA complex. We set out to determine the structure and mechanism of this pentameric complex, based on this published data. However, we were not able to reconstitute a pentameric complex following the procedures described. Instead, we could unambiguously show that, in our hands, addition of TAF11/TAF13 rather dissociates TFIIA/TBP/DNA. We demonstrated that TAF11/TAF13 displaces DNA from the trimeric complex, and sequesters TBP resulting in a hitherto unknown, stable TAF11/TAF13/TBP. We characterized this novel complex by biophysical and biochemical analyses, and could demonstrate by native MS and AUC that a stoichiometry of 1:1:1 in TAF11/TAF13/TBP. We provide evidence that in this TAF11/TAF13/TBP complex,

TAF11/TAF13 binds to the DNA binding surface of TBP. This surface is incidentally also the surface, TBP uses as a homodimerization interface. The binding of TAF11/TAF13 to the concave DNA binding interface of TBP is reminiscent of the interactions found between the N-terminal domain of TAF1 (TAF1-TAND) and TBP, an observation which was termed 'TATA-box mimicry'. It is thought that obstructing the DNA binding interface of TBP by TAF1 in TFIID may play a role in transcriptional regulation at the core promoter. Moreover, obstruction of the DNA binding surface in TBP was also found in a complex between the transcriptional regulator Mot1 and TBP. We have shown that TAF11/TAF13 binding to TBP and TAF1-TAND binding to TBP is mutually exclusive, arguing for TAF11/TAF13 and TAF1-TAND sharing the same binding site on the concave interface of TBP. It is interesting to speculate how this finding relates to TFIID assembly in the cell. Our results suggest that only TAF1-TAND or TAF11/TAF13 can interact with TBP within TFIID, and either of these interactions will be disrupted when TFIID binds to the core promoter via TBP. Our results provide a structural framework for further studies on TAF11/TAF13/TBP, and we intend to follow up with the aim to obtain interaction information at atomic resolution. This data can then be used to probe the functional consequences of TAF11/TAF13 binding to TBP in *in vivo*.

4.3. Towards structural characterization of 9TAF complex

Our current protocol in the Berger lab to produce fully recombinant and active human holo-TFIID relies on a reconstitution step *in vitro* where a highly purified 9TAF complex, containing TAF2, TAF3, TAF4, TAF5, TAF6, TAF8, TAF9, TAF10 and TAF12 is reacted with an MBPTAF1-Module which contains the remaining TAF1, TAF7, TAF11, TAF13 and TBP subunits of TFIID. We speculate that this *in vitro* reconstitution step may relate to mechanisms of TFIID assembly which may also occur *in vivo*, although further experiments are clearly required to establish this. Structure analysis of the 9TAF complex can provide

information about how holo-TFIID is crafted. In the third part of my thesis, I have carried out structural analysis of highly purified mildly cross-linked recombinant 9TAF complex by negative-stain electron microscopy and prepared structural models by the random conical tilt (RCT) method. This work resulted in calculation of four RCT models, indicating the presence of conformational heterogeneity in the sample. Interestingly, one of the RCT models already adopts a shape reminiscent of clamp shaped holo-TFIID determined by EM studies of endogenous sample purified from native source material. My results set the stage for high resolution cryo-EM structure determination, as the models I obtained can serve as starting models for particle sorting during cryo-EM data analysis. A high resolution cryo-EM structure, capitalizing on recent developments in EM structure determination technologies and softwares (the cryo-EM revolution) will be instrumental to decipher the inner working of TFIID and mechanisms of holo-TFIID assembly.

CONCLUSION AND OUTLOOK

The results presented in this thesis provide important new insights into molecular interactions at the core of eukaryotic transcription initiation and they also raise many interesting new questions. A broad range of structural, biochemical and biophysical experimental tools and technologies were brought to bear to obtain these results, illustrating the potential of an integrated approach to elucidate complex functional architectures. The results presented set the stage for further studies, to correlate the structural findings to function *in vivo*. Moreover and importantly, these findings also enable us now to significantly improve the production protocols we have in place in the laboratory for producing fully recombinant holo-TFIID, unlocking this essential General Transcription Factor for high resolution structural and mechanistic analyses, *in vitro* and *in vivo*, in the near future.

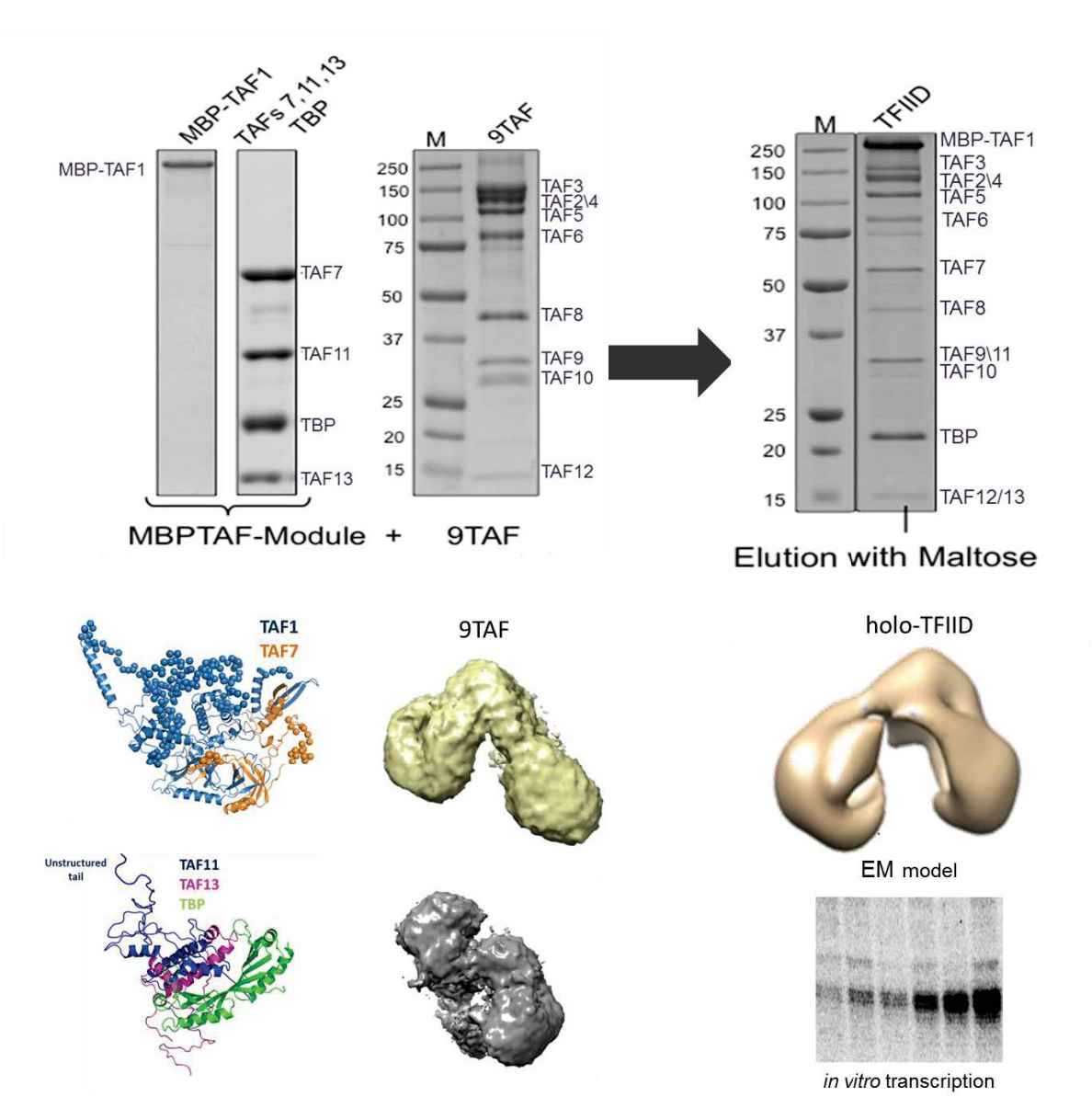
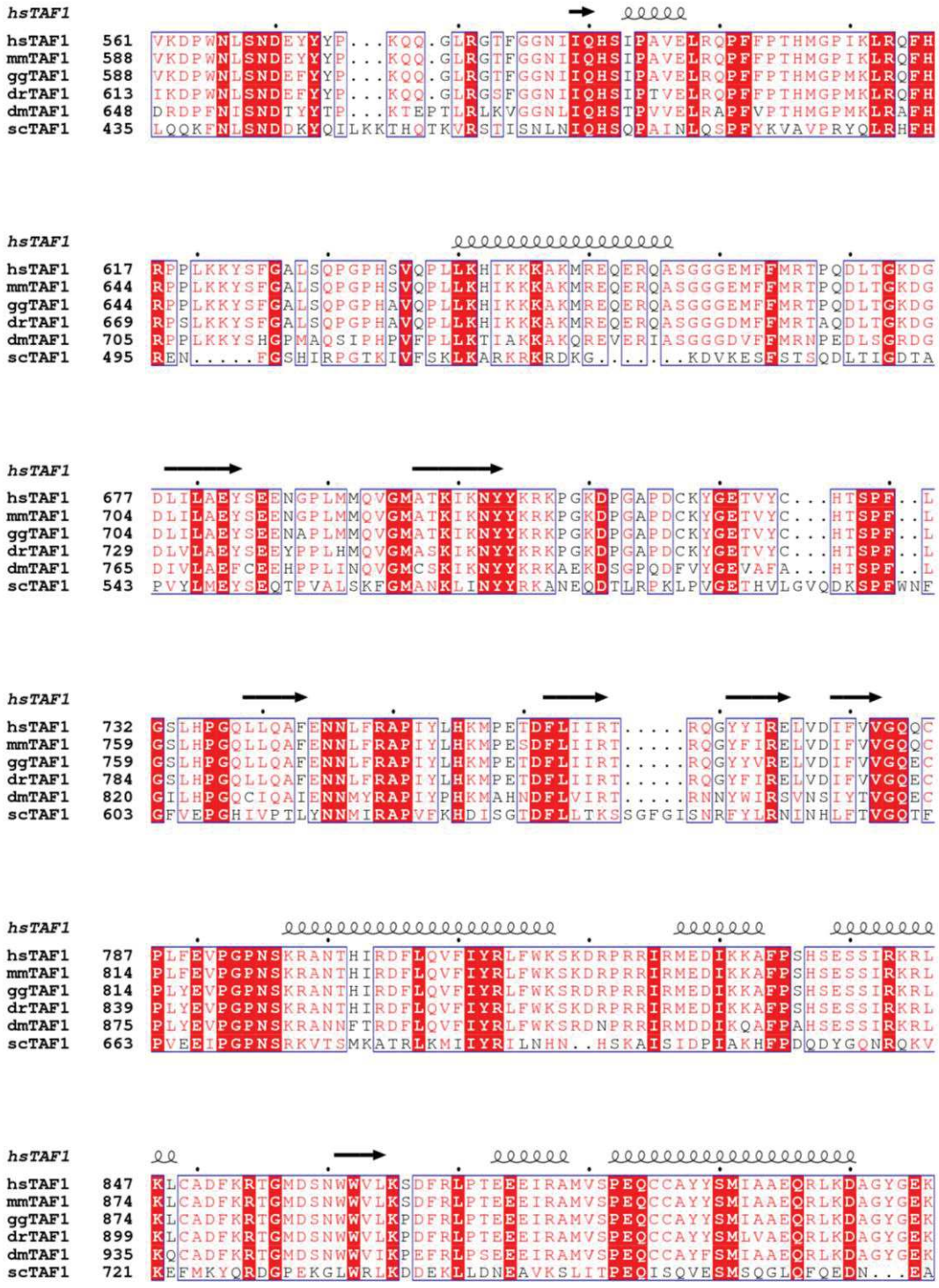


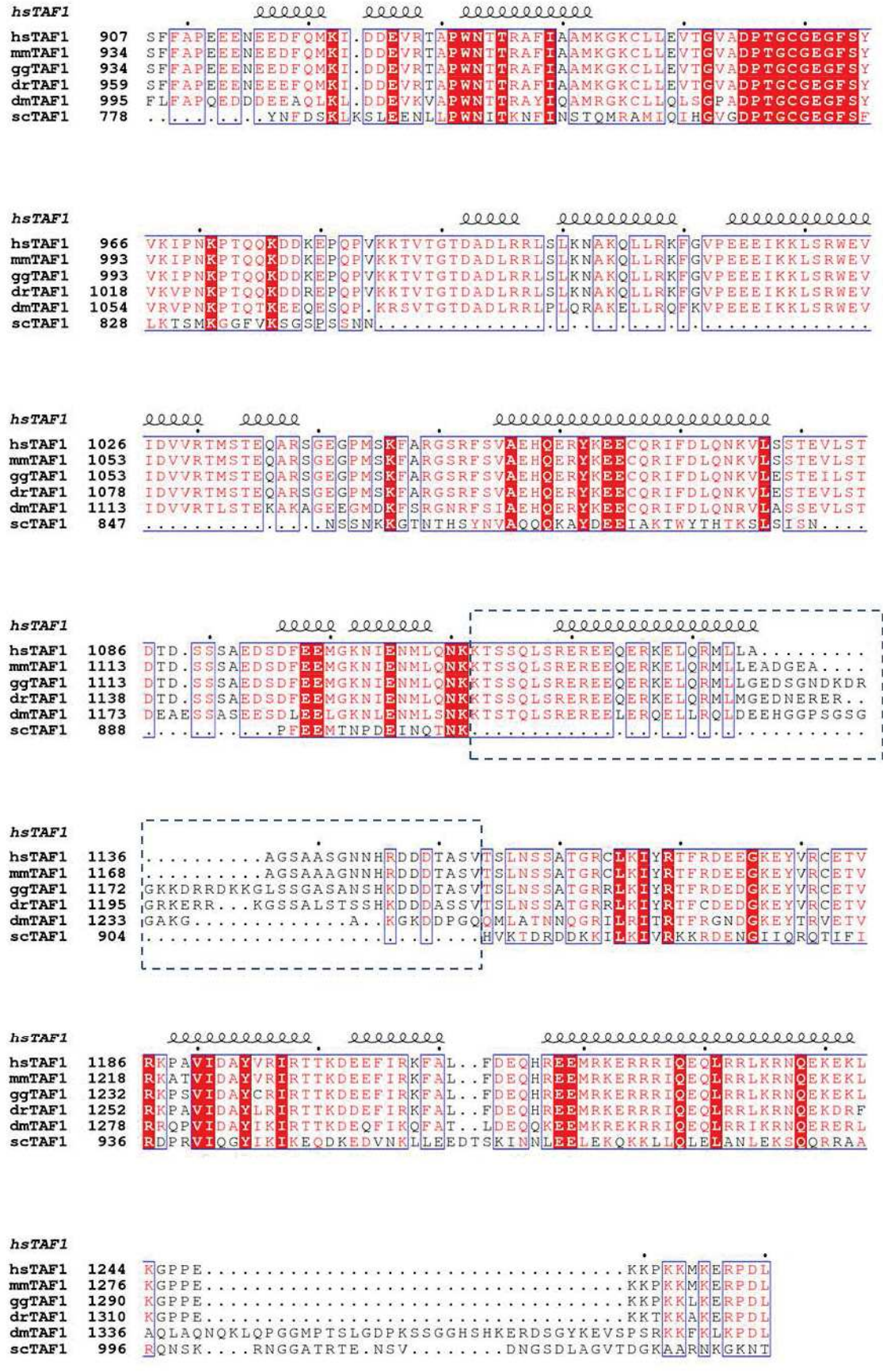
Figure 29: Summary of the studies on human TFIID subassemblies (this work) in holo-TFIID context.

Procedure for reconstitution of holo-TFIID in Berger lab is shown in top panel. First a MBPTAF-Module is mixed with preformed 9TAF complex to reconstitute holo-TFIID. SDS-PAGE analysis is shown here. This recombinant TFIID adopts a horse-shoe shape as shown by preliminary EM and is active *in vitro* transcription assay (below). Structural analysis of TFIID subassemblies performed in this work is depicted. A complete structure of TAF1/TAF7 complex, a quasi-atomic model of novel TAF11/TAF13/TBP complex and two most prominent negative stain EM models of 9TAF complex are shown.

5. APPENDIX AND SUPPLEMENTS

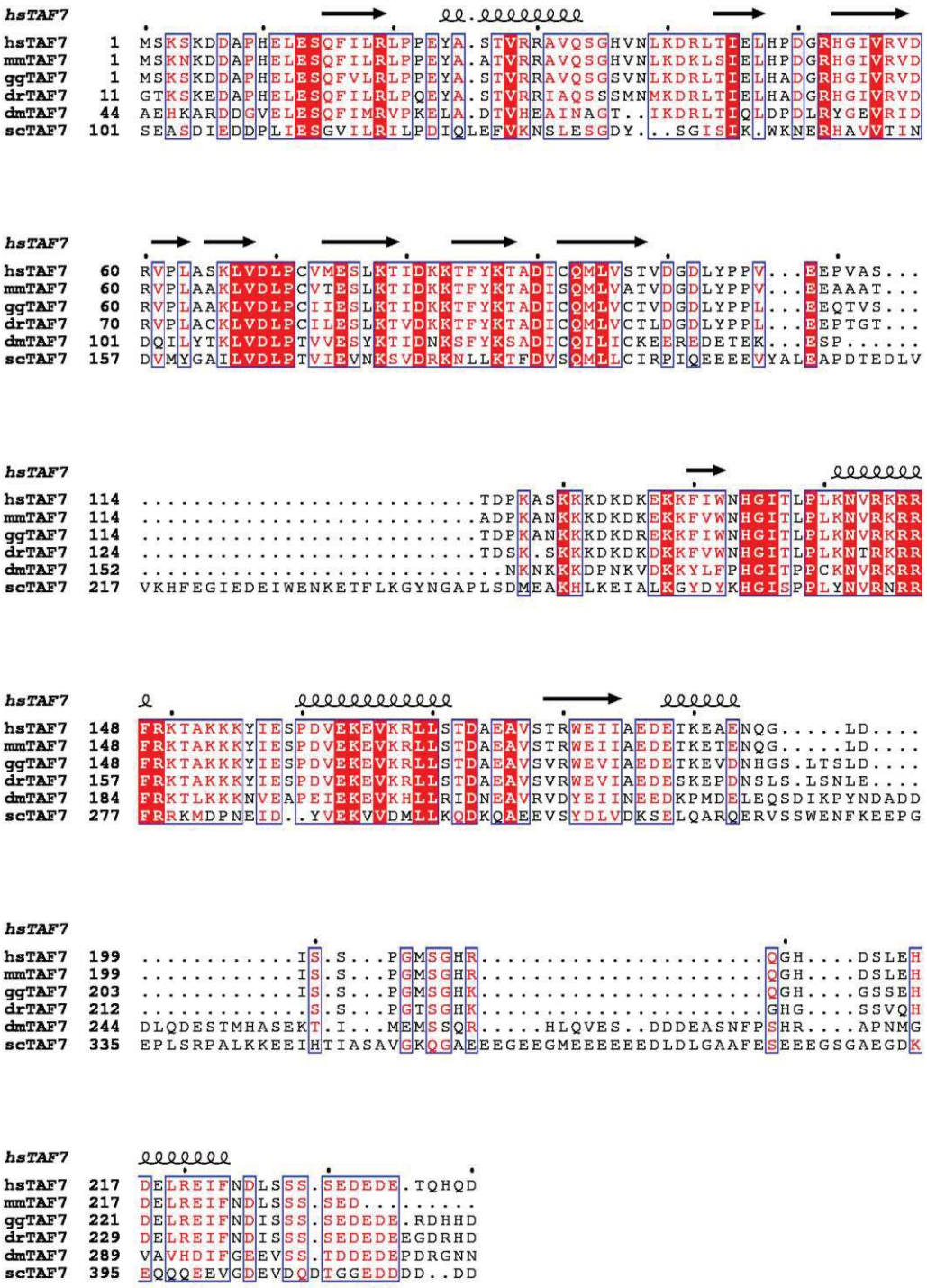
5.1. Multispecies sequence alignment of TAF1




hsTAF1 907 SFFAPEEENEEDFQMKI.DDEVRTAPWNTTRAFIAAMKGKCLLEVTGVADPTGCGEGFSY
mmTAF1 934 SFFAPEEENEEDFQMKI.DDEVRTAPWNTTRAFIAAMKGKCLLEVTGVADPTGCGEGFSY
ggTAF1 934 SFFAPEEENEEDFQMKI.DDEVRTAPWNTTRAFIAAMKGKCLLEVTGVADPTGCGEGFSY
drTAF1 959 SFFAPEEENEEEFQMKI.DDEVRTAPWNTTRAFIAAMKGKCLLEVTGVADPTGCGEGFSY
dmTAF1 995 FLFAPQEDDDEEAQLKL.DDEVKVAPWNTTRAYIQAMRGKCLLQLSGPADPTGCGEGFSY
scTAF1 778 ..YNFDSKLKSLEENLLPWNITKNFINSTQMRAMIQIHGVGDPTGCGEGFSF
hsTAF1 966 VKIPNKPTQQKDDKEPQPVKKTVTGTDADLRRLSLKNAKQLLRKFGVPEEEIKKLSRWEV
mmTAF1 993 VKIPNKPTQQKDDKEPQPVKKTVTGTDADLRRLSLKNAKQLLRKFGVPEEEIKKLSRWEV
ggTAF1 993 VKIPNKPTQQKDDKEPQPVKKTVTGTDADLRRLSLKNAKQLLRKFGVPEEEIKKLSRWEV
drTAF1 1018 VKVPNKPTQQKDDREPQPVKKTVTGTDADLRRLSLKNAKQLLRKFGVPEEEIKKLSRWEV
dmTAF1 1054 VRVPNKPTQTKEEQESQP.KRSVTGTDADLRRLPLQRAKELLRQFKVPEEEIKKLSRWEV
scTAF1 828 LKTSMKGGFVKSGSPSSNN
hsTAF1 1026 IDVVRTMSTEQARSGEGPMSKFARGSRFSVAEHQERYKEECQRIFDLQNKVLSSTEVLST
mmTAF1 1053 IDVVRTMSTEQARSGEGPMSKFARGSRFSVAEHQERYKEECQRIFDLQNKVLSSTEVLST
ggTAF1 1053 IDVVRTMSTEQARSGEGPMSKFARGSRFSVAEHQERYKEECQRIFDLQNKVLESTEILST
drTAF1 1078 IDVVRTMSTEQARSGEGPMSKFARGSRFSVAEHQERYKEECQRIFDLQNKVLESTEVLST
dmTAF1 1113 IDVVRTLSTEKAKAGEEGMDKFSRGNRFSIAEHQERYKEECQRIFDLQNRVLASSEVLST
scTAF1 847 .....NSSNKKGTNTHSYNVAQQQKAYDEEIAKTWYTHTKSLSISN....
hsTAF1 1086 DTD.SSSAEDSDFEEMGKNIENMLQNKKTSSQLSREREEQERKELQRMLLA.....
mmTAF1 1113 DTD.SSSAEDSDFEEMGKNIENMLQNKKTSSQLSREREEQERKELQRMLLEADGEA....
ggTAF1 1113 DTD.SSSAEDSDFEEMGKNIENMLQNKKTSSQLSREREEQERKELQRMLLGEDSGNDKDR
drTAF1 1138 DTD.SSSAEDSDFEEMGKNIENMLQNKKTSSQLSREREEQERKELQRMLMGEDNERER..
dmTAF1 1173 DEAESSSASEESDLEELGKNLENMLSNKKTSTQLSREREELERQELLRQLDEEHGGPSGSG
scTAF1 888 ...PFEEMTNPDEINQTNKL
hsTAF1 1136 .....AGSAASGNNHRDDDTASVTSLNSSATGRCLKIYRTFRDEEGKEYVRCETV
mmTAF1 1168 .....AGSAAAGNNHRDDDTASVTSLNSSATGRCLKIYRTFRDEEGKEYVRCETV
ggTAF1 1172 GKKDRRDKKGLSSGASANSHKDDDTASVTSLNSSATGRRLKIYRTFRDEDGKEYVRCETV
drTAF1 1195 GRKERR...KGSSALSTSSHKDDDASSVTSLNSSATGRRLKIYRTFCDEDGKEYVRCETV
dmTAF1 1233 GAKG.....A..KGKDDPGQQMLATNNQGRILRITRTFRGNDGKEYTRVETV
scTAF1 904 .....HVKTDRDDKKILKIVRKKRDENGIIQRQTIFI
hsTAF1 1186 RKPAVIDAYVRIRTTKDEEFIRKFALFDEQHREEMRKERRRIQEQLRRLKRNQEKEKL
mmTAF1 1218 RKATVIDAYVRIRTTKDEEFIRKFALFDEQHREEMRKERRRIQEQLRRLKRNQEKEKL
ggTAF1 1232 RKPSVIDAYCRIRTTKDEEFIRKFALFDEQHREEMRKERRRIQEQLRRLKRNQEKEKL
drTAF1 1252 RKPAVIDAYLRIRTTKDDEFIRKFALFDEQHREEMRKERRRIQEQLRRLKRNQEKDRF
dmTAF1 1278 RRQPVIDAYIKIRTTKDEQFIKQFATLDEQQKEEMKREKRRIQEQLRRIKRNQERERL
scTAF1 936 RDPRVIQGYIKIKEQDKEDVNKLLEEDTSKINNLEELEKQKKLLQLELANLEKSQQRRAA
hsTAF1 1244 KGPPE.....KKPKKMKERPDL
mmTAF1 1276 KGPPE.....KKPKKMKERPDL
ggTAF1 1290 KGPPE.....KKPKKLKERPDL
drTAF1 1310 KGPPE.....KKTKKAKERPDL
dmTAF1 1336 AQLAQNQKLQPGGMPTSLGDPKSSGGHSHKERDSGYKEVSPSRKKFKLKPDL
scTAF1 996 RQNSK....RNGGATRTE.NSV.....DNGSDLAGVTDGKAARNKGKNT

Multispecies Sequence alignment of TAF1 is shown only for human residue 561-1260. The partially structured 50AA acid stretch inserted in human is shown with dotted box. [***ClustalOmega*** *(http://www.ebi.ac.uk/Tools/msa/clustalo/) (McWilliam et al., 2013; Sievers and Higgins, 2014) and* ***ESPRIPT*** *(http://espript.ibcp.fr/ESPript/ESPript/) (Gouet et al., 1999)]*

5.2. Multispecies sequence alignment of TAF7



Multispecies Sequence alignment of TAF1 is shown only for human residue 1-240. [***ClustalOmega*** *(http://www.ebi.ac.uk/Tools/msa/clustalo/) (McWilliam et al., 2013; Sievers and Higgins, 2014) and* ***ESPRIPT*** *(http://espript.ibcp.fr/ESPript/ESPript/) (Gouet et al., 1999)]*

5.3. Intrinsic disorder propensity for yeast and human TAF1

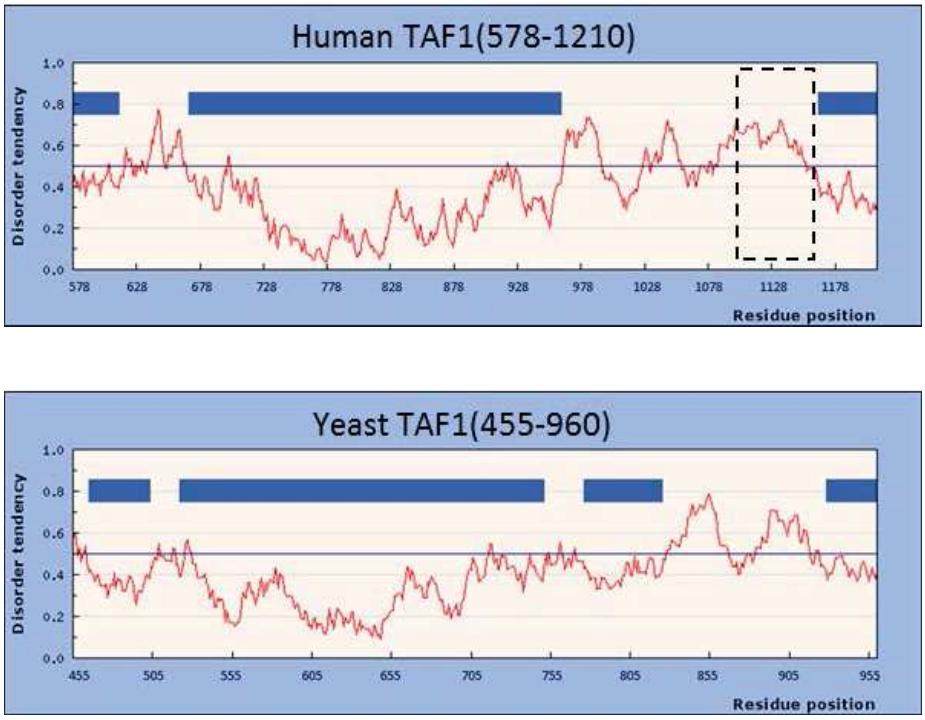
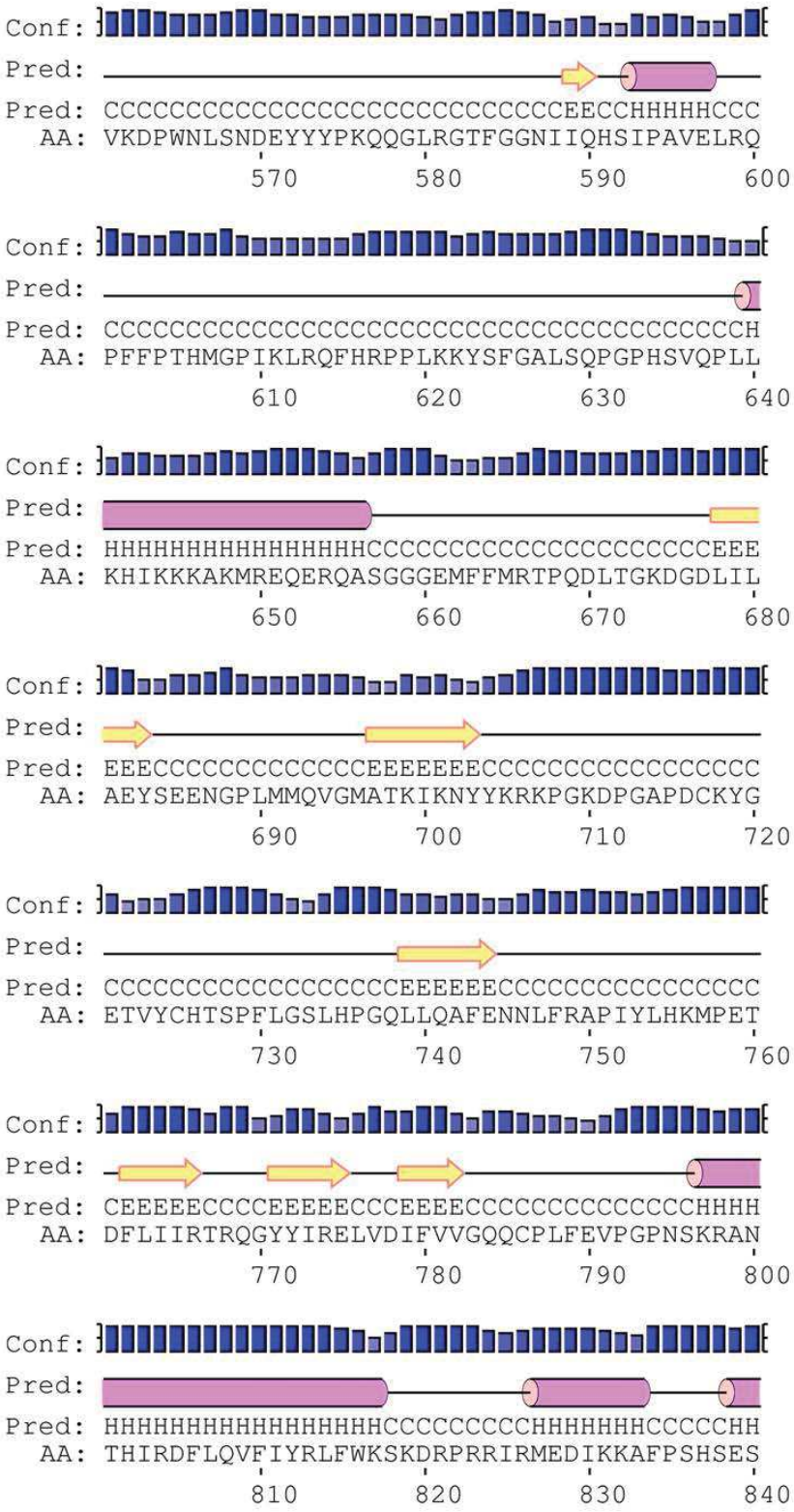
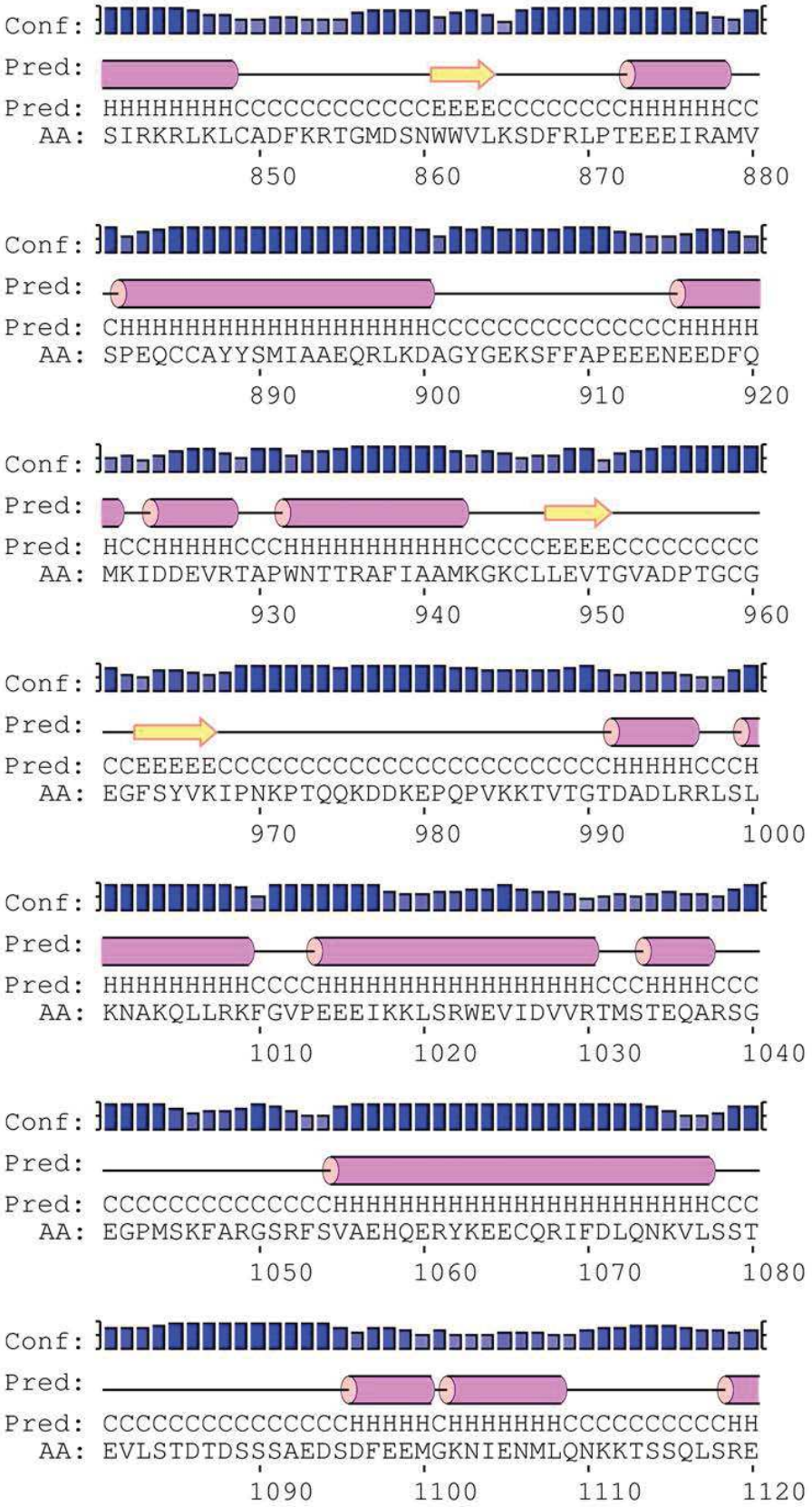


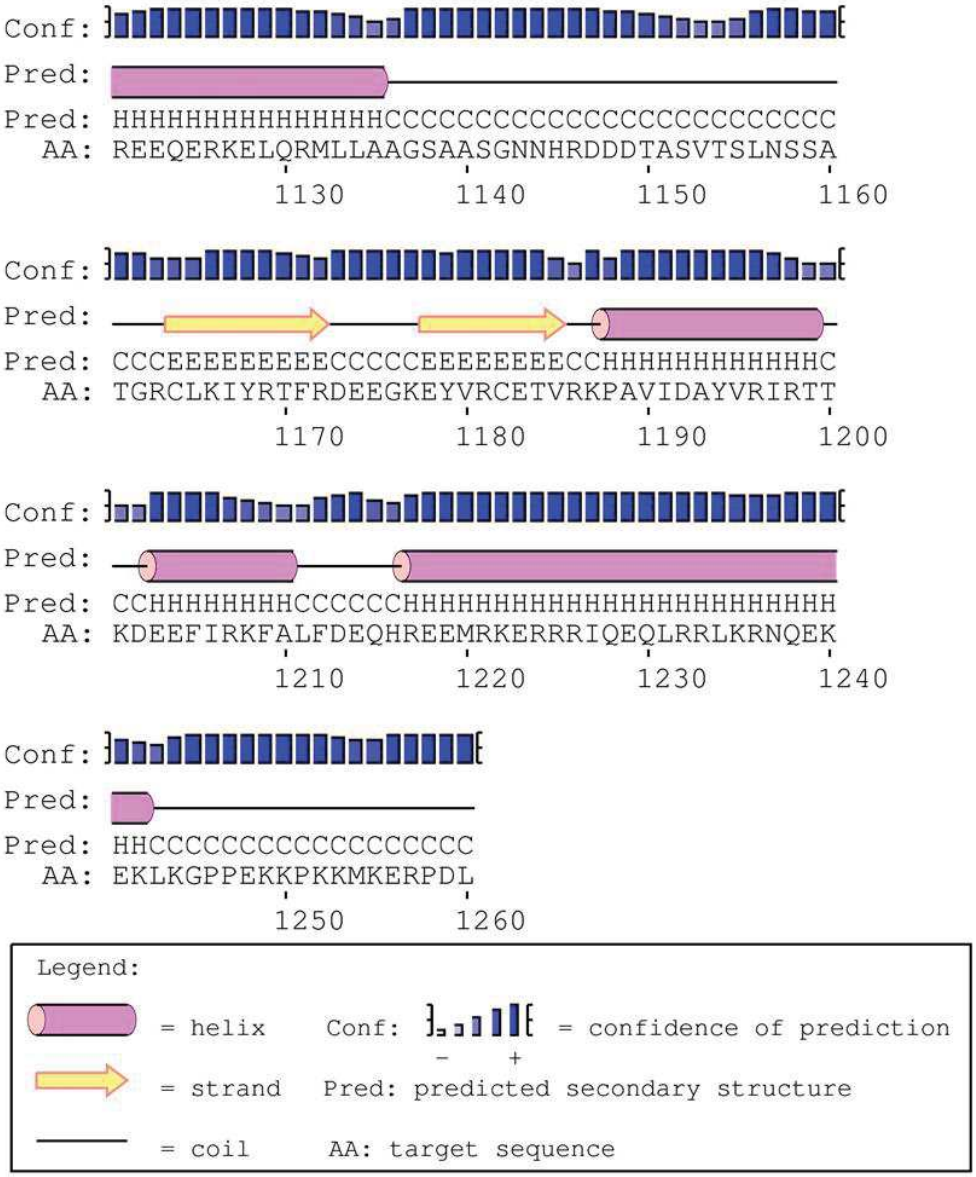
Figure 30: Intrinsic disorder propensity for TAF1 from Human and Yeast.

Disorder propensity is shown on Y-axis for each residue (X-axis). The 50AA acid stretch inserted in human is shown with dotted box which has a higher disorder propensity. ***[IUPRED*** *(http://iupred.enzim.hu) (Dosztanyi et al., 2005)]*

5.4. Secondary structure prediction of human TAF1



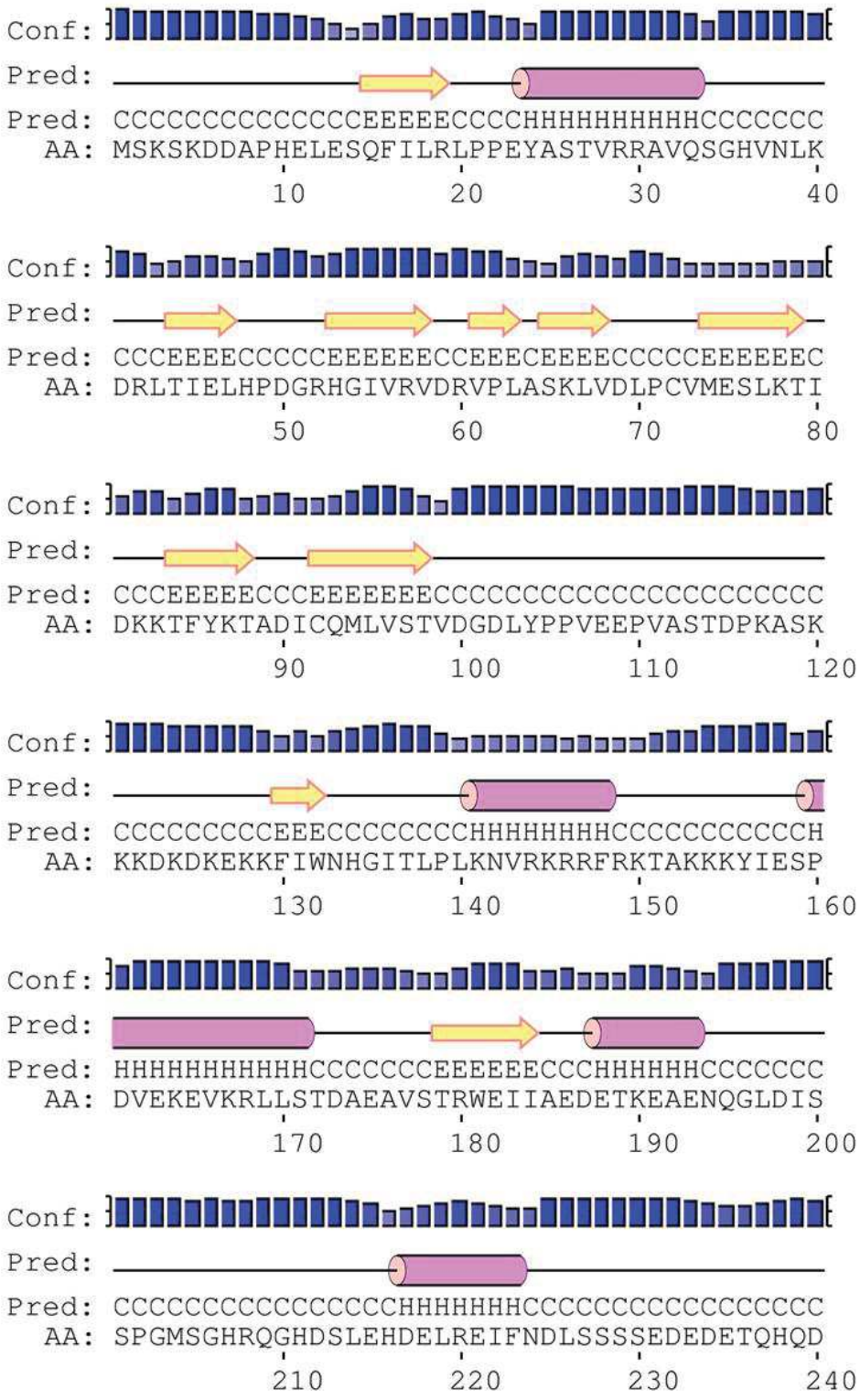

Conf:
Pred:
Pred: HHHHHHHHCCCCCCCCCCCCEEEECCCCCCCCHHHHHHCC
AA: SIRKRLKLCADFKRTGMDSNWWVLKSDFRLPTEEEIRAMV
850 860 870 880
Conf:
Pred:
Pred: CHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCHHHHH
AA: SPEQCCAYYSMIAAEQRLKDAGYGEKSFFAPEEENEEDFQ
890 900 910 920
Conf:
Pred:
Pred: HCCHHHHHCCCHHHHHHHHHHCCCCCEEEECCCCCCCCC
AA: MKIDDEVRTAPWNTTRAFIAAMKGKCLLEVTGVADPTGCG
930 940 950 960
Conf:
Pred:
Pred: CCEEEEECCCCCCCCCCCCCCCCCCCCCCCHHHHHCCCH
AA: EGFSYVKIPNKPTQQKDDKEPQPVKKTVTGTDADLRRLSL
970 980 990 1000
Conf:
Pred:
Pred: HHHHHHHHHCCCCHHHHHHHHHHHHHHHHCCCHHHHCCC
AA: KNAKQLLRKFGVPEEEIKKLSRWEVIDVVRTMSTEQARSG
1010 1020 1030 1040
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCC
AA: EGPMSKFARGSRFSVAEHQERYKEECQRIFDLQNKVLSST
1050 1060 1070 1080
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCHHHHHCHHHHHHHCCCCCCCCCHH
AA: EVLSTDTDSSSAEDSDFEEMGKNIENMLQNKKTSSQLSRE
1090 1100 1110 1120



Secondary structure prediction of human TAF1 is shown only for residue 561-1260.

*[**PSIPRED** (http://bioinf.cs.ucl.ac.uk/psipred) (Jones, 1999)]*

5.5. Secondary structure prediction of human TAF7



Secondary structure prediction of human TAF7 is shown only for residue 1-240. Legend is similar to as for secondary structure prediction for human TAF1. *[**PSIPRED** (http://bioinf.cs.ucl.ac.uk/psipred) (Jones, 1999)]*

5.6. Matthew's coefficient calculation for crystallization data of TAF1$^{578-1106}$/TAF7$^{1-193}$

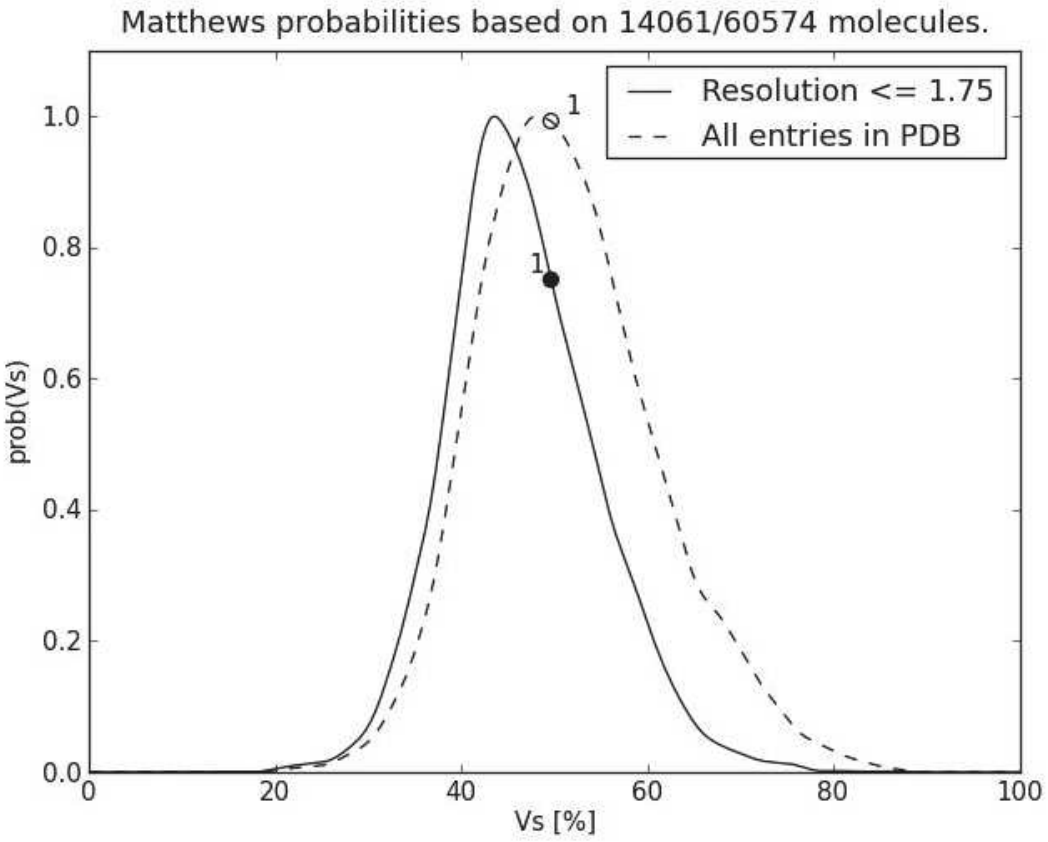


Figure 31: Matthew's probabilities plot for crystal data of TAF1$^{578-1106}$/TAF7$^{1-193}$.

The percentage solvent content corresponding to the number of molecules is shown on the X-axis and the corresponding probability is shown on the Y-axis.

Table 12: Matthew's probabilities for crystallization data of TAF1$^{578-1106}$/TAF7$^{1-193}$.

The table lists the data corresponding to Figure 31.

Number of molecules	Probability for resolution	Probability overall	V_m (Å^3/Da)	Solvent (%)	MW (Da)
1	1.0	1.0	2.44	49.61	82996

5.7. Molprobity validation of refined structure of TAF1$^{578-1106}$/TAF7$^{1-193}$

All-Atom Contacts	Clashscore, all atoms:	1.5		99th percentile* (N=932, 1.750Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	2	0.42%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	510	97.89%	Goal: >98%
	MolProbity score^	0.92		100th percentile* (N=11713, 1.750Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad backbone bonds:	0 / 4426	0.00%	Goal: 0%
	Bad backbone angles:	1 / 5963	0.02%	Goal: <0.1%

In the two column results, the left column gives the raw count, right column gives the percentage.
* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.
^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

Figure 32: Molprobity analysis of the refined structure of TAF1$^{578-1106}$/TAF7$^{1-193}$.

Stereochemistry of the refined structure is good, similar to the best structures of similar resolution as indicated by a low clashscore (1.5) and low Molprobity score (0.92).

5.8. Kratky plot, P(r), Guinier analysis and structural parameters for SAXS data of TAF1/TAF7

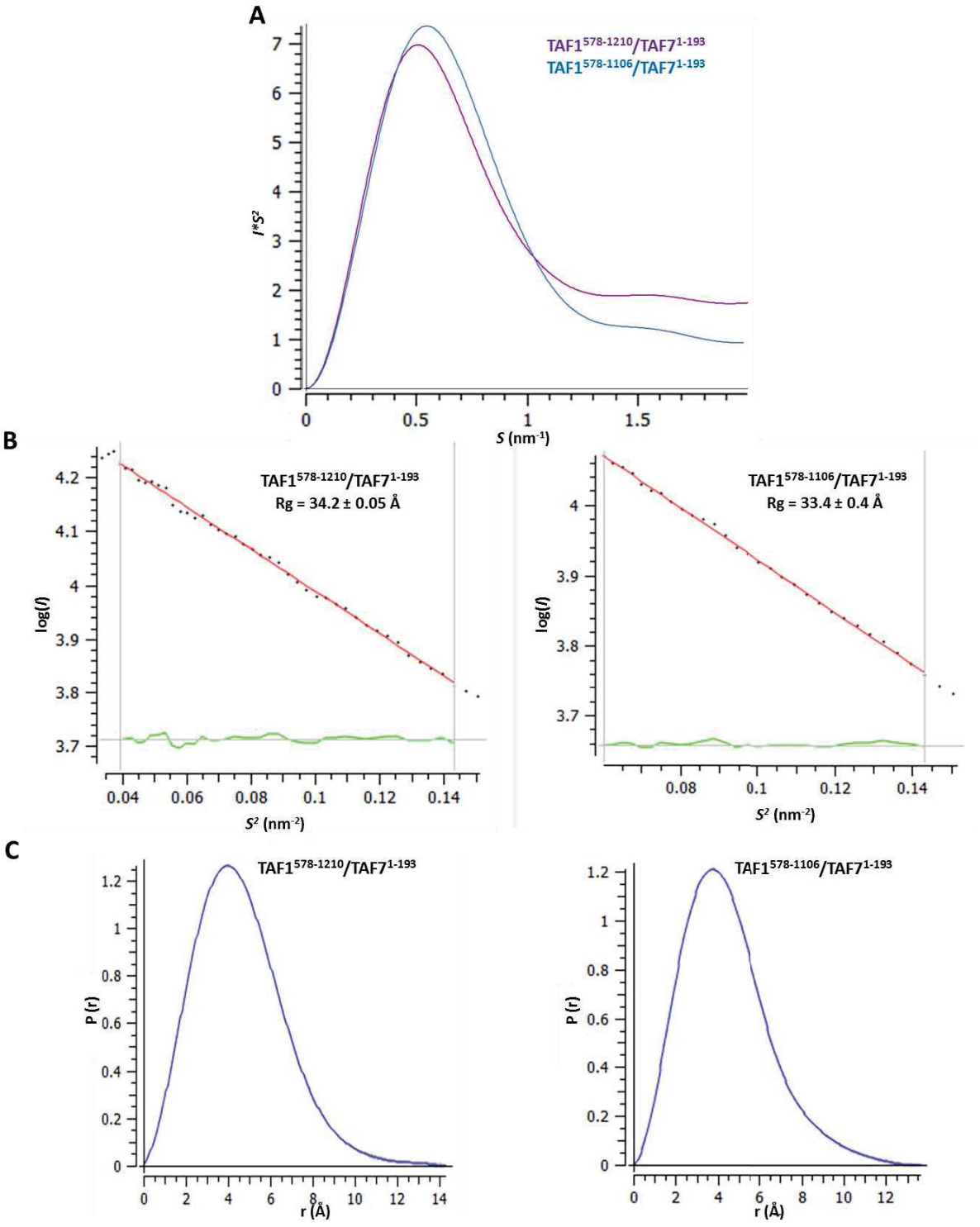


Figure 33: Kratky plot, P(r) and Guinier analysis for SAXS data of TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$.

(A) Kratky plots are shown for TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ with magenta and light blue colors respectively, indicating globular protein with some flexibility. **(B)** Guinier plots are shown for TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$ along with corresponding Rg values, indicating a straight line matching with data points used for Rg calculation by Guinier. Green lines show the residuals of the linear fit. **(C)** Distance distribution analysis for TAF1$^{578-1210}$/TAF7$^{1-193}$ (left) and TAF1$^{578-1106}$/TAF7$^{1-193}$ (right).

Table 13: Data collection and structural parameters for SAXS analysis of TAF1$^{578-1210}$/TAF7$^{1-193}$ and TAF1$^{578-1106}$/TAF7$^{1-193}$.

	TAF1$^{578-1210}$/TAF7$^{1-193}$	TAF1$^{578-1106}$/TAF7$^{1-193}$
Data collection parameters		
Beamline	ESRF-BM29	ESRF-BM29
Beam size at sample	~700 µm x 700 µm	~700 µm x 700 µm
Wavelength (Å)	0.9919	0.9919
S range (Å^{-1})	0.003-0.497	0.003-0.497
Concentration range (mg ml^{-1})	0.3-3.03	0.36-3.07
Temperature	4°C	4°C
Structural parameters†		
I(0) (arbitrary units) [from *P*(*r*)]	81.32	74.53
Rg (Å) [from *P*(*r*)]	35.5	30.46
I(0) (arbitrary units) (from Guinier)	79.86 ± 0.24	73.28 ± 0.17
Rg (Å) (from Guinier)	34.2 ± 0.05	33.4 ± 0.4
D_{max} (Å)	142.9	136.6
Porod volume estimate (Å^3)	185250	167250
Molecular mass M_r [from porod volume]	108.97 kDa	98.38 kDa

† Reported for experimental merged data.

5.9. SEC-MALLS of TAF1$^{1157-1207}$/ TAF7$^{153-190}$

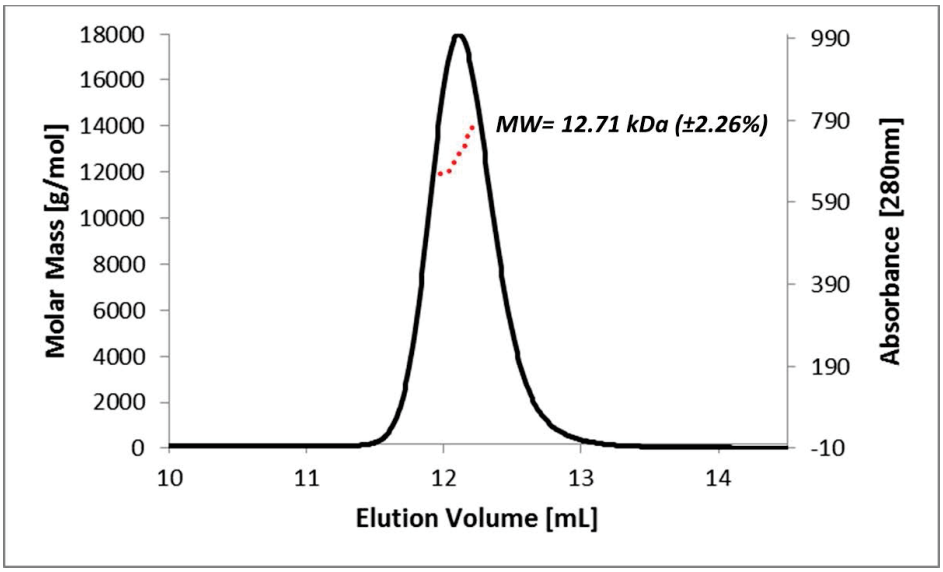


Figure 34: SEC-MALLS of TAF1$^{1157-1207}$/ TAF7$^{153-190}$

SEC-MALLS of TAF1$^{1157-1207}$/ TAF7$^{153-190}$ shows the presence of a single species in the sample with a molecular weight of ~12.71 kDa (expected MW= 12.032 kDa), which corresponds to the monomer of heterodimer.

5.10. Intrinsic disorder propensity for human TAF11 and TAF13

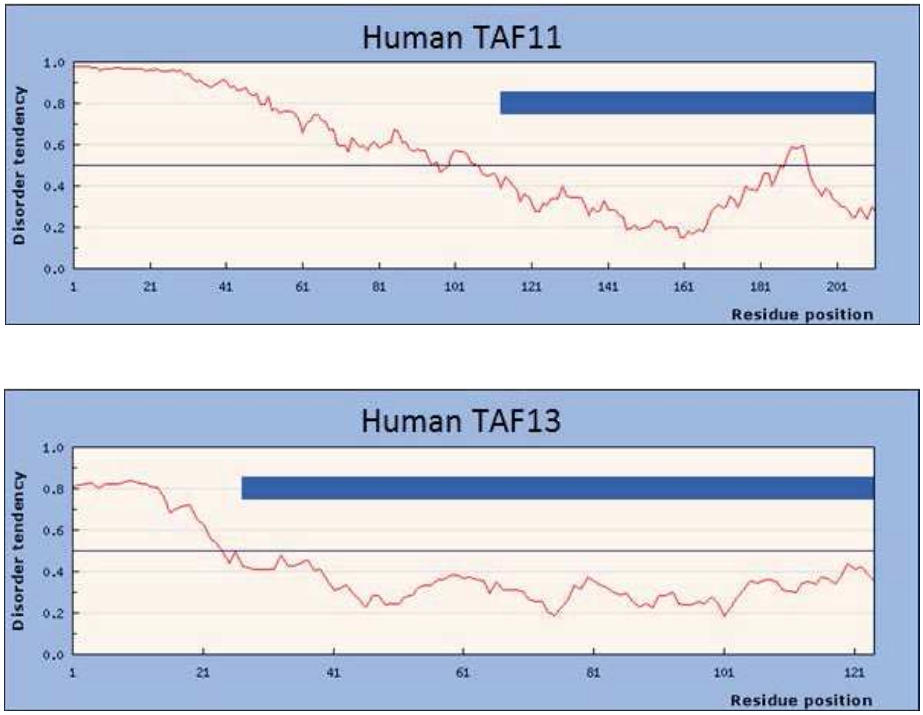


Figure 35: Intrinsic disorder propensity for human TAF11 and TAF13.

Disorder propensity is shown on Y-axis for each residue (X-axis). The 50AA acid stretch inserted in human is shown with dotted box which has a higher disorder propensity. ***[IUPRED*** *(http://iupred.enzim.hu) (Dosztanyi et al., 2005)]*

5.11. Different constructs of human TAF11/TAF13 complex

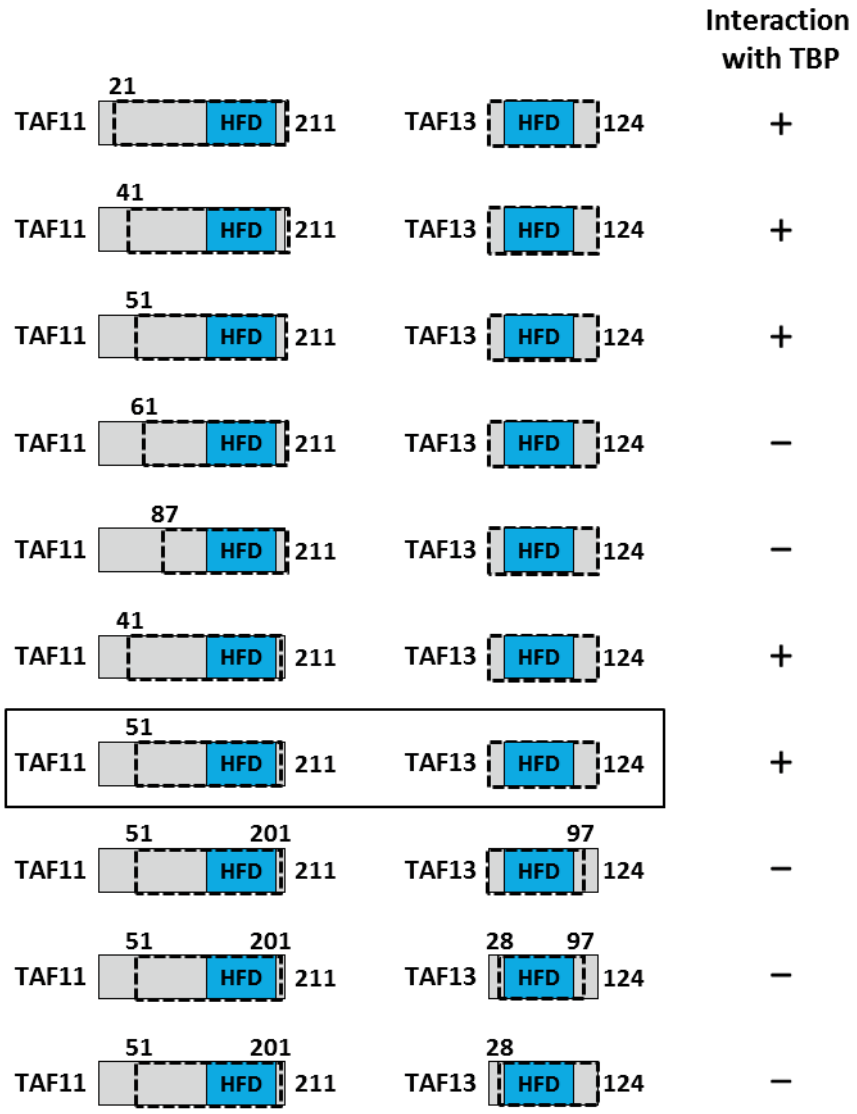


Figure 36: Different constructs of human TAF11/TAF13 complex.

Different construct of TAF11/TAF13 are shown with dotted lines. Interaction of each construct with core domain of human TBP is also represented as determined by SEC analysis.

5.12. Kratky plot, P(r), Guinier analysis and structural parameters for SAXS data of TAF11/TAF13/TBP and TAF11/TAF13

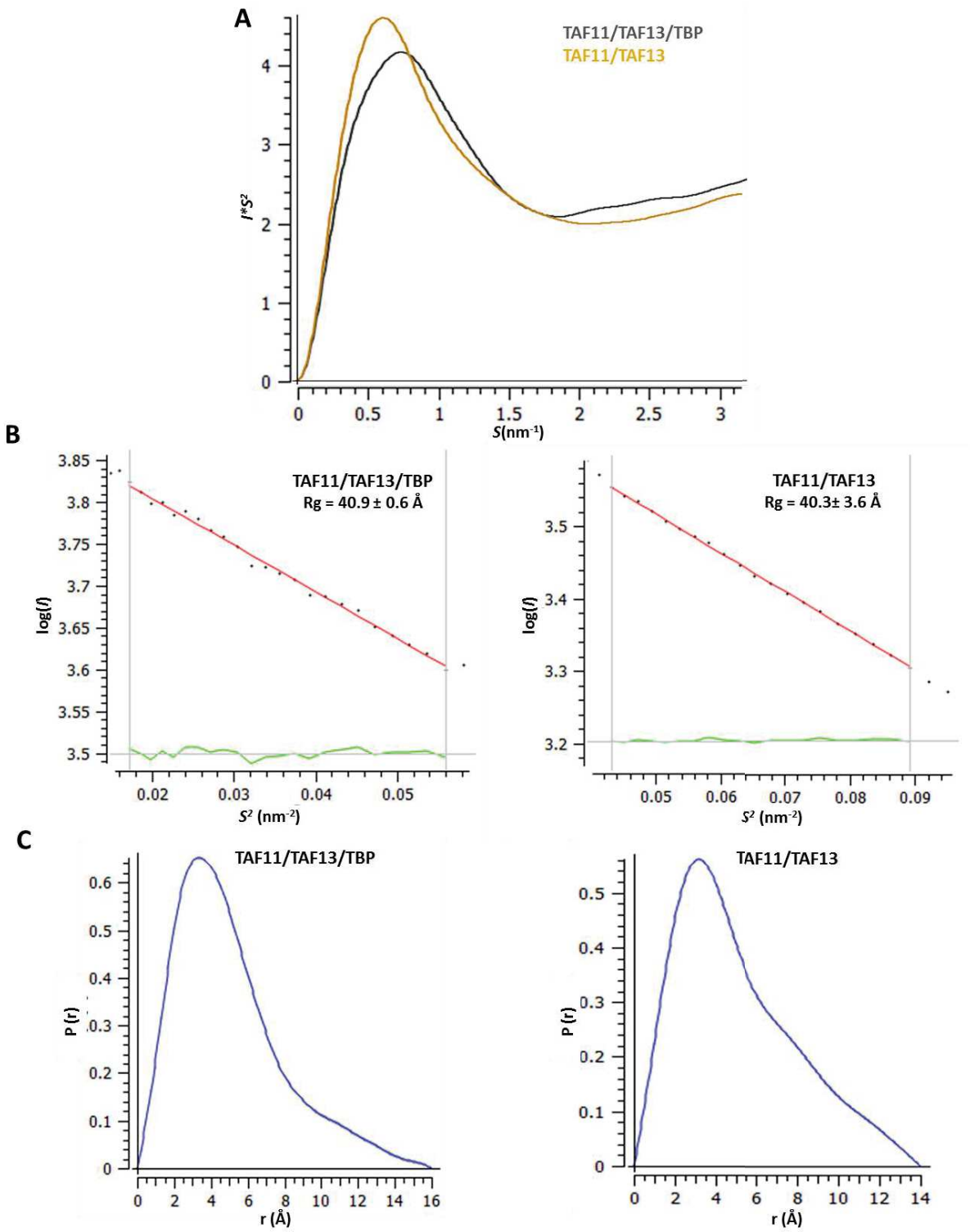


Figure 37: Kratky plot, P(r) and Guinier analysis for SAXS data of TAF11/TAF13/TBP and TAF11/TAF13.

(A) Kratky plots are shown for TAF11/TAF13/TBP and TAF11/TAF13 with light grey and light orange colors respectively, indicating presence of flexibility with globular domains. **(B)** Guinier plots are shown for TAF11/TAF13/TBP and TAF11/TAF13 along with corresponding Rg values, indicating a straight line matching with data points used for Rg calculation by Guinier. Green lines show the residuals of the linear fit. **(C)** Distance distribution analysis for TAF11/TAF13/TBP (left) and TAF11/TAF13 (right).

Table 14: Data collection and structural parameters for SAXS analysis of TAF11/TAF13/TBP and TAF11/TAF13.

	TAF11/TAF13/TBP	**TAF11/TAF13**
Data collection parameters		
Beamline	ESRF-BM29	ESRF-BM29
Beam size at sample	~700 µm x 700 µm	~700 µm x 700 µm
Wavelength (Å)	0.931	0.931
S range ($Å^{-1}$)	0.003-0.497	0.003-0.497
Concentration range (mg ml^{-1})	0.3-7.11	0.53-7.48
Temperature (°C)	4°C	4°C
Structural parameters†		
I(0) (arbitrary units) [from *P*(*r*)]	49.63	43.65
Rg (Å) [from *P*(*r*)]	41.0	41.2
I(0) (arbitrary units) (from Guinier)	50.21 ± 0.33	44.15 ± 0.08
Rg (Å) (from Guinier)	40.9 ± 0.6	40.3± 3.6
D_{max} (Å)	160	140
Porod volume estimate ($Å^3$)	120110	89850
Molecular mass M_r [from porod volume]	70.65 kDa	52.91 kDa

† Reported for experimental merged data.

5.13. Cross-links in TAF11/TAF13/TBP complex by CLMS after BS3 crosslinking.

Table 15: Crosslinks present in BS3 cross-linked TAF11/TAF13/TBP complex.

TAF11 - TBP			
TAF11 Residue	**TBP Residue**	**No. of matches**	**Highest Score**
K97	K293	5	9.416
K85	K293	2	8.818
K82	K239	2	8.314
K135	K191	1	8.085
K82	K293	2	7.926
K82	K191	1	7.684
K131	K293	1	6.928
K83	K232	1	6.124
K95	K293	1	4.145
TAF13 - TBP			
TAF13 Residue	**TBP Residue**	**No. of matches**	**Highest Score**
K34	K177	1	10.578
K34	K191	1	7.856
K101	K232	1	7.584
K115	K191	1	7.519
K115	K177	1	5.742
TAF11 - TAF13			
TAF11 Residue	**TAF13 Residue**	**No. of matches**	**Highest Score**
K195	K96	2	12.795
K85	K101	2	12.014
K204	K96	3	11.053
K89	K101	5	10.98
K204	K101	6	10.909
K131	K34	3	9.627
K82	K92	2	9.321
K197	K96	2	9.308
K131	K111	10	9.226
K135	K111	2	8.371
K131	K115	1	8.326
K207	K92	1	8.292
K82	K96	1	7.981
K85	K115	1	7.943

K94	K101	2	7.934
K97	K101	1	7.867
K204	K92	4	6.872
K85	K92	1	6.817
K82	K101	1	6.437
K206	K92	1	5.77
K204	S74	1	3.42
K206	S74	1	3.42

REFERENCES

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica Section D, Biological crystallography *66*, 213-221.

Anandapadamanaban, M., Andresen, C., Helander, S., Ohyama, Y., Siponen, M.I., Lundstrom, P., Kokubo, T., Ikura, M., Moche, M., and Sunnerhagen, M. (2013). High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. Nature structural & molecular biology *20*, 1008-1014.

Andel, F., 3rd, Ladurner, A.G., Inouye, C., Tjian, R., and Nogales, E. (1999). Three-dimensional structure of the human TFIID-IIA-IIB complex. Science *286*, 2153-2156.

Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E., and Moudrianakis, E.N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proceedings of the National Academy of Sciences of the United States of America *88*, 10148-10152.

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics *22*, 195-201.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature *226*, 1209-1211.

Bell, B., Scheer, E., and Tora, L. (2001). Identification of hTAF(II)80 delta links apoptotic signaling pathways to transcription factor TFIID function. Molecular cell *8*, 591-600.

Berger, I., Fitzgerald, D.J., and Richmond, T.J. (2004). Baculovirus expression system for heterologous multiprotein complexes. Nature biotechnology *22*, 1583-1587.

Bhattacharya, S., Lou, X., Hwang, P., Rajashankar, K.R., Wang, X., Gustafsson, J.A., Fletterick, R.J., Jacobson, R.H., and Webb, P. (2014). Structural and functional insight into TAF1-TAF7, a subcomplex of

transcription factor II D. Proceedings of the National Academy of Sciences of the United States of America *111*, 9103-9108.

Bhattacharya, S., Takada, S., and Jacobson, R.H. (2007). Structural analysis and dimerization potential of the human TAF5 subunit of TFIID. Proceedings of the National Academy of Sciences of the United States of America *104*, 1189-1194.

Bieniossek, C., Imasaki, T., Takagi, Y., and Berger, I. (2012). MultiBac: expanding the research toolbox for multiprotein complexes. Trends in biochemical sciences *37*, 49-57.

Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O.*, et al.* (2009). Automated unrestricted multigene recombineering for multiprotein complex production. Nat Methods *6*, 447-450.

Bieniossek, C., Papai, G., Schaffitzel, C., Garzoni, F., Chaillet, M., Scheer, E., Papadopoulos, P., Tora, L., Schultz, P., and Berger, I. (2013). The architecture of human general transcription factor TFIID core complex. Nature *493*, 699-702.

Bieniossek, C., Richmond, T.J., and Berger, I. (2008). MultiBac: multigene baculovirus-based eukaryotic protein complex production. Current protocols in protein science / editorial board, John E Coligan [et al] *Chapter 5*, Unit 5 20.

Birck, C., Poch, O., Romier, C., Ruff, M., Mengus, G., Lavigne, A.C., Davidson, I., and Moras, D. (1998). Human TAF(II)28 and TAF(II)18 interact through a histone fold encoded by atypical evolutionary conserved motifs also found in the SPT3 family. Cell *94*, 239-249.

Bleichenbacher, M., Tan, S., and Richmond, T.J. (2003). Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. J Mol Biol *332*, 783-793.

Brand, M., Leurent, C., Mallouh, V., Tora, L., and Schultz, P. (1999). Three-dimensional structures of the TAFII-containing complexes TFIID and TFTC. Science *286*, 2151-2153.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S.*, et al.* (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta crystallographica Section D, Biological crystallography *54*, 905-921.

Buratowski, S., Hahn, S., Sharp, P.A., and Guarente, L. (1988). Function of a yeast TATA element-binding protein in a mammalian transcription system. Nature *334*, 37-42.

Burke, T.W., and Kadonaga, J.T. (1997). The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes & development *11*, 3020-3031.

Cang, Y., and Prelich, G. (2002). Direct stimulation of transcription by negative cofactor 2 (NC2) through TATA-binding protein (TBP). Proceedings of the National Academy of Sciences of the United States of America *99*, 12727-12732.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C.*, et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics *38*, 626-635.

Cavallini, B., Huet, J., Plassat, J.L., Sentenac, A., Egly, J.M., and Chambon, P. (1988). A yeast activity can substitute for the HeLa cell TATA box factor. Nature *334*, 77-80.

Chalkley, G.E., and Verrijzer, C.P. (1999). DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. The EMBO journal *18*, 4835-4845.

Chamberlin, M., and Berg, P. (1962). Deoxyribo ucleic acid-directed synthesis of ribonucleic acid by an enzyme from Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America *48*, 81-94.

Chasman, D.I., Flaherty, K.M., Sharp, P.A., and Kornberg, R.D. (1993). Crystal structure of yeast TATA-binding protein and model for interaction with DNA. Proceedings of the National Academy of Sciences of the United States of America *90*, 8174-8178.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta crystallographica Section D, Biological crystallography *66*, 12-21.

Chiang, C.M., and Roeder, R.G. (1995). Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. Science *267*, 531-536.

Cianfrocco, M.A., Kassavetis, G.A., Grob, P., Fang, J., Juven-Gershon, T., Kadonaga, J.T., and Nogales, E. (2013). Human TFIID binds to core promoter DNA in a reorganized structural state. Cell *152*, 120-131.

Cler, E., Papai, G., Schultz, P., and Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor TFIID. Cellular and molecular life sciences : CMLS *66*, 2123-2134.

Coleman, R.A., Taggart, A.K., Burma, S., Chicca, J.J., 2nd, and Pugh, B.F. (1999). TFIIA regulates TBP and TFIID dimers. Molecular cell *4*, 451-457.

Cormack, B.P., Strubin, M., Ponticelli, A.S., and Struhl, K. (1991). Functional differences between yeast and human TFIID are localized to the highly conserved region. Cell *65*, 341-348.

Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138-163.

Crowley, T.E., Hoey, T., Liu, J.K., Jan, Y.N., Jan, L.Y., and Tjian, R. (1993). A new factor related to TATA-binding protein has highly restricted expression patterns in Drosophila. Nature *361*, 557-561.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 3rd, Snoeyink, J., Richardson, J.S.*, et al.* (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic acids research *35*, W375-383.

de Sanctis, D., Beteva, A., Caserotto, H., Dobias, F., Gabadinho, J., Giraud, T., Gobbo, A., Guijarro, M., Lentini, M., Lavault, B.*, et al.* (2012). ID29: a high-intensity highly automated ESRF beamline for macromolecular crystallography experiments exploiting anomalous scattering. Journal of synchrotron radiation *19*, 455-461.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. Journal of biomolecular NMR *6*, 277-293.

Demeny, M.A., Soutoglou, E., Nagy, Z., Scheer, E., Janoshazi, A., Richardot, M., Argentini, M., Kessler, P., and Tora, L. (2007). Identification of a small TAF complex and its role in the assembly of TAF-containing complexes. PloS one *2*, e316.

Deng, W., and Roberts, S.G. (2005). A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes & development *19*, 2418-2423.

Dikstein, R., Zhou, S., and Tjian, R. (1996). Human TAFII 105 is a cell type-specific TFIID subunit related to hTAFII130. Cell *87*, 137-146.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Dvir, A., Tan, S., Conaway, J.W., and Conaway, R.C. (1997). Promoter escape by RNA polymerase II. Formation of an escape-competent transcriptional intermediate is a prerequisite for exit of polymerase from the promoter. The Journal of biological chemistry *272*, 28175-28178.

Dynlacht, B.D., Hoey, T., and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell *66*, 563-576.

Elmlund, H., Baraznenok, V., Linder, T., Szilagyi, Z., Rofougaran, R., Hofer, A., Hebert, H., Lindahl, M., and Gustafsson, C.M. (2009). Cryo-EM reveals promoter DNA binding and conformational flexibility of the general transcription factor TFIID. Structure *17*, 1442-1452.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta crystallographica Section D, Biological crystallography *60*, 2126-2132.

Falender, A.E., Freiman, R.N., Geles, K.G., Lo, K.C., Hwang, K., Lamb, D.J., Morris, P.L., Tjian, R., and Richards, J.S. (2005a). Maintenance of spermatogenesis requires TAF4b, a gonad-specific subunit of TFIID. Genes & development *19*, 794-803.

Falender, A.E., Shimada, M., Lo, Y.K., and Richards, J.S. (2005b). TAF4b, a TBP associated factor, is required for oocyte development and function. Developmental biology *288*, 405-419.

Filippakopoulos, P., Picaud, S., Fedorov, O., Keller, M., Wrobel, M., Morgenstern, O., Bracher, F., and Knapp, S. (2012). Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. Bioorganic & medicinal chemistry *20*, 1878-1886.

Fishburn, J., and Hahn, S. (2012). Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. Molecular and cellular biology *32*, 12-25.

Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., and Berger, I. (2006). Protein complex expression by using multigene baculoviral vectors. Nat Methods *3*, 1021-1032.

Flot, D., Mairs, T., Giraud, T., Guijarro, M., Lesourd, M., Rey, V., van Brussel, D., Morawe, C., Borel, C., Hignette, O.*, et al.* (2010). The ID23-2 structural biology microfocus beamline at the ESRF. Journal of synchrotron radiation *17*, 107-118.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. Journal of applied crystallography *42*, 342-346.

Freiman, R.N., Albright, S.R., Zheng, S., Sha, W.C., Hammer, R.E., and Tjian, R. (2001). Requirement of tissue-selective TBP-associated factor TAFII105 in ovarian development. Science *293*, 2084-2087.

Frontini, M., Soutoglou, E., Argentini, M., Bole-Feysot, C., Jost, B., Scheer, E., and Tora, L. (2005). TAF9b (formerly TAF9L) is a bona fide TAF that has unique and overlapping roles with TAF9. Molecular and cellular biology *25*, 4638-4649.

Gabadinho, J., Beteva, A., Guijarro, M., Rey-Bakaikoa, V., Spruce, D., Bowler, M.W., Brockhauser, S., Flot, D., Gordon, E.J., Hall, D.R.*, et al.* (2010). MxCuBE: a synchrotron beamline control environment customized for macromolecular crystallography experiments. Journal of synchrotron radiation *17*, 700-707.

Gaiser, F., Tan, S., and Richmond, T.J. (2000). Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 A resolution. J Mol Biol *302*, 1119-1127.

Gazit, K., Moshonov, S., Elfakess, R., Sharon, M., Mengus, G., Davidson, I., and Dikstein, R. (2009). TAF4/4b x TAF12 displays a unique mode of DNA binding and is required for core promoter function of a subset of genes. The Journal of biological chemistry *284*, 26286-26296.

Gegonne, A., Devaiah, B.N., and Singer, D.S. (2013). TAF7: traffic controller in transcription initiation. Transcription *4*, 29-33.

Gegonne, A., Weissman, J.D., and Singer, D.S. (2001). TAFII55 binding to TAFII250 inhibits its acetyltransferase activity. Proceedings of the National Academy of Sciences of the United States of America *98*, 12432-12437.

Gegonne, A., Weissman, J.D., Zhou, M., Brady, J.N., and Singer, D.S. (2006). TAF7: a possible transcription initiation check-point regulator. Proceedings of the National Academy of Sciences of the United States of America *103*, 602-607.

Georgieva, S., Kirschner, D.B., Jagla, T., Nabirochkina, E., Hanke, S., Schenkel, H., de Lorenzo, C., Sinha, P., Jagla, K., Mechler, B., *et al.* (2000). Two novel Drosophila TAF(II)s have homology with human TAF(II)30 and are differentially regulated during development. Molecular and cellular biology *20*, 1639-1648.

Gershenzon, N.I., and Ioshikhes, I.P. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics *21*, 1295-1300.

Gill, G., Pascal, E., Tseng, Z.H., and Tjian, R. (1994). A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. Proceedings of the National Academy of Sciences of the United States of America *91*, 192-196.

Gouet, P., Courcelle, E., Stuart, D.I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics *15*, 305-308.

Gowri, P.M., Yu, J.H., Shaufl, A., Sperling, M.A., and Menon, R.K. (2003). Recruitment of a repressosome complex at the growth hormone receptor promoter and its potential role in diabetic nephropathy. Molecular and cellular biology *23*, 815-825.

Grewal, S.I., and Jia, S. (2007). Heterochromatin revisited. Nature reviews Genetics *8*, 35-46.

Grob, P., Cruse, M.J., Inouye, C., Peris, M., Penczek, P.A., Tjian, R., and Nogales, E. (2006). Cryo-electron microscopy studies of human TFIID: conformational breathing in the integration of gene regulatory cues. Structure *14*, 511-520.

Groft, C.M., Uljon, S.N., Wang, R., and Werner, M.H. (1998). Structural homology between the Rap30 DNA-binding domain and linker histone H5:

implications for preinitiation complex assembly. Proceedings of the National Academy of Sciences of the United States of America *95*, 9117-9122.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nature structural & molecular biology *11*, 394-403.

Hahn, S., Buratowski, S., Sharp, P.A., and Guarente, L. (1989). Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. Proceedings of the National Academy of Sciences of the United States of America *86*, 5718-5722.

Han, Y., Luo, J., Ranish, J., and Hahn, S. (2014). Architecture of the Saccharomyces cerevisiae SAGA transcription coactivator complex. The EMBO journal *33*, 2534-2546.

Hansen, S.K., Takada, S., Jacobson, R.H., Lis, J.T., and Tjian, R. (1997). Transcription properties of a cell type-specific TATA-binding protein, TRF. Cell *91*, 71-83.

He, Y., Fang, J., Taatjes, D.J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. Nature *495*, 481-486.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA polymerase IV directs silencing of endogenous DNA. Science *308*, 118-120.

Herrera, F.J., Yamaguchi, T., Roelink, H., and Tjian, R. (2014). Core promoter factor TAF9B regulates neuronal gene expression. eLife *3*, e02559.

Heymann, J.B., and Belnap, D.M. (2007). Bsoft: image processing and molecular modeling for electron microscopy. J Struct Biol *157*, 3-18.

Hilton, T.L., Li, Y., Dunphy, E.L., and Wang, E.H. (2005). TAF1 histone acetyltransferase activity in Sp1 activation of the cyclin D1 promoter. Molecular and cellular biology *25*, 4321-4332.

Hoffmann, A., Chiang, C.M., Oelgeschlager, T., Xie, X., Burley, S.K., Nakatani, Y., and Roeder, R.G. (1996). A histone octamer-like structure within TFIID. Nature *380*, 356-359.

Holstege, F.C., Fiedler, U., and Timmers, H.T. (1997). Three transitions in the RNA polymerase II transcription complex during initiation. The EMBO journal *16*, 7468-7480.

Horn, P.J., and Peterson, C.L. (2002). Molecular biology. Chromatin higher order folding--wrapping up transcription. Science *297*, 1824-1827.

Huisinga, K.L., Brower-Toland, B., and Elgin, S.C. (2006). The contradictory definitions of heterochromatin: transcription and silencing. Chromosoma *115*, 110-122.

Hurwitz, J., Bresler, A., and Diringer, R. (1960). The enzymic incorporation of ribonucleotides into polyribonucleotides and the effect of DNA. Biochemical and biophysical research communications *3*, 15-19.

Huth, J.R., Bewley, C.A., Jackson, B.M., Hinnebusch, A.G., Clore, G.M., and Gronenborn, A.M. (1997). Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. Protein science : a publication of the Protein Society *6*, 2359-2364.

Imhof, A., Yang, X.J., Ogryzko, V.V., Nakatani, Y., Wolffe, A.P., and Ge, H. (1997). Acetylation of general transcription factors by histone acetyltransferases. Current biology : CB *7*, 689-692.

Jacobson, R.H., Ladurner, A.G., King, D.S., and Tjian, R. (2000). Structure and function of a human TAFII250 double bromodomain module. Science *288*, 1422-1425.

Jacq, X., Brou, C., Lutz, Y., Davidson, I., Chambon, P., and Tora, L. (1994). Human TAFII30 is present in a distinct TFIID complex and is required for transcriptional activation by the estrogen receptor. Cell *79*, 107-117.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol *292*, 195-202.

Jung, Y.S., and Zweckstetter, M. (2004). Mars -- robust automatic backbone assignment of proteins. Journal of biomolecular NMR *30*, 11-23.

Juven-Gershon, T., and Kadonaga, J.T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. Developmental biology *339*, 225-229.

Kabsch, W. (2010). Xds. Acta crystallographica Section D, Biological crystallography *66*, 125-132.

Kadonaga, J.T. (2002). The DPE, a core promoter element for transcription by RNA polymerase II. Experimental & molecular medicine *34*, 259-264.

Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R.G., Meisterernst, M., and Burley, S.K. (2001). Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. Cell *106*, 71-81.

Kambadur, R., Culotta, V., and Hamer, D. (1990). Cloned yeast and mammalian transcription factor TFIID gene products support basal but not activated metallothionein gene transcription. Proceedings of the National Academy of Sciences of the United States of America *87*, 9168-9172.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nature genetics *37*, 761-765.

Kanno, T., Kanno, Y., Siegel, R.M., Jang, M.K., Lenardo, M.J., and Ozato, K. (2004). Selective recognition of acetylated histones by bromodomain proteins visualized in living cells. Molecular cell *13*, 33-43.

Kastner, B., Fischer, N., Golas, M.M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E.*, et al.* (2008). GraFix: sample preparation for single-particle electron cryomicroscopy. Nat Methods *5*, 53-55.

Kaufmann, J., and Smale, S.T. (1994). Direct recognition of initiator elements by a component of the transcription factor IID complex. Genes & development *8*, 821-829.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature *365*, 520-527.

Kim, T.K., Ebright, R.H., and Reinberg, D. (2000). Mechanism of ATP-dependent promoter melting by transcription factor IIH. Science *288*, 1418-1422.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512-520.

Kim, Y.J., Bjorklund, S., Li, Y., Sayre, M.H., and Kornberg, R.D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell *77*, 599-608.

Kloet, S.L., Whiting, J.L., Gafken, P., Ranish, J., and Wang, E.H. (2012). Phosphorylation-dependent regulation of cyclin D1 and cyclin A gene transcription by TFIID subunits TAF1 and TAF7. Molecular and cellular biology *32*, 3358-3369.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. Journal of applied crystallography *36*, 1277-1282.

Korkhin, Y., Unligil, U.M., Littlefield, O., Nelson, P.J., Stuart, D.I., Sigler, P.B., Bell, S.D., and Abrescia, N.G. (2009). Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure. PLoS biology *7*, e1000102.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868-871.

Kornberg, R.D. (1999). Eukaryotic transcriptional control. Trends in cell biology *9*, M46-49.

Kornberg, R.D., and Thomas, J.O. (1974). Chromatin structure; oligomers of the histones. Science *184*, 865-868.

Kraemer, S.M., Ranallo, R.T., Ogg, R.C., and Stargell, L.A. (2001). TFIIA interacts with TFIID via association with TATA-binding protein and TAF40. Molecular and cellular biology *21*, 1737-1746.

Kutach, A.K., and Kadonaga, J.T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Molecular and cellular biology *20*, 4754-4764.

Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., and Ebright, R.H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes & development *12*, 34-44.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. Journal of biomolecular NMR *8*, 477-486.

LaVallie, E.R., DiBlasio-Smith, E.A., Collins-Racie, L.A., Lu, Z., and McCoy, J.M. (2003). Thioredoxin and related proteins as multifunctional fusion tags for soluble expression in E. coli. Methods Mol Biol *205*, 119-140.

Leal, R.M., Bourenkov, G.P., Svensson, O., Spruce, D., Guijarro, M., and Popov, A.N. (2011). Experimental procedure for the characterization of radiation damage in macromolecular crystals. Journal of synchrotron radiation *18*, 381-386.

Lee, D.H., Gershenzon, N., Gupta, M., Ioshikhes, I.P., Reinberg, D., and Lewis, B.A. (2005). Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. Molecular and cellular biology *25*, 9674-9686.

Lescop, E., Schanda, P., and Brutscher, B. (2007). A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson *187*, 163-169.

Leurent, C., Sanders, S., Ruhlmann, C., Mallouh, V., Weil, P.A., Kirschner, D.B., Tora, L., and Schultz, P. (2002). Mapping histone fold TAFs within yeast TFIID. The EMBO journal *21*, 3424-3433.

Leurent, C., Sanders, S.L., Demeny, M.A., Garbett, K.A., Ruhlmann, C., Weil, P.A., Tora, L., and Schultz, P. (2004). Mapping key functional sites within yeast TFIID. The EMBO journal *23*, 719-727.

Li, M.Z., and Elledge, S.J. (2007). Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods *4*, 251-256.

Lim, C.Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., and Kadonaga, J.T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. Genes & development *18*, 1606-1617.

Littlefield, O., Korkhin, Y., and Sigler, P.B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. Proceedings of the National Academy of Sciences of the United States of America *96*, 13668-13673.

Liu, D., Ishima, R., Tong, K.I., Bagby, S., Kokubo, T., Muhandiram, D.R., Kay, L.E., Nakatani, Y., and Ikura, M. (1998). Solution structure of a TBP-TAF(II)230 complex: protein mimicry of the minor groove surface of the TATA box unwound by TBP. Cell *94*, 573-583.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Luse, D.S., and Jacob, G.A. (1987). Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter. The Journal of biological chemistry *262*, 14990-14997.

Maile, T., Kwoczynski, S., Katzenberger, R.J., Wassarman, D.A., and Sauer, F. (2004). TAF1 activates transcription by phosphorylation of serine 33 in histone H2B. Science *304*, 1010-1014.

Mal, T.K., Masutomi, Y., Zheng, L., Nakata, Y., Ohta, H., Nakatani, Y., Kokubo, T., and Ikura, M. (2004). Structural and functional characterization on the interaction of yeast TFIID subunit TAF1 with TATA-binding protein. J Mol Biol *339*, 681-693.

Malik, S., Guermah, M., and Roeder, R.G. (1998). A dynamic model for PC4 coactivator function in RNA polymerase II transcription. Proceedings of the National Academy of Sciences of the United States of America *95*, 2192-2197.

Malik, S., Lee, D.K., and Roeder, R.G. (1993). Potential RNA polymerase II-induced interactions of transcription factor TFIIB. Molecular and cellular biology *13*, 6253-6259.

Malkowska, M., Kokoszynska, K., Rychlewski, L., and Wyrwicz, L. (2013). Structural bioinformatics of the general transcription factor TFIID. Biochimie *95*, 680-691.

Marsh, J.A., Singh, V.K., Jia, Z., and Forman-Kay, J.D. (2006). Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein science : a publication of the Protein Society *15*, 2795-2804.

Martinez, E. (2002). Multi-protein complexes in eukaryotic gene transcription. Plant molecular biology *50*, 925-947.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annual review of genomics and human genetics *7*, 29-59.

Maston, G.A., Zhu, L.J., Chamberlain, L., Lin, L., Fang, M., and Green, M.R. (2012). Non-canonical TAF complexes regulate active promoters in human embryonic stem cells. eLife *1*, e00068.

Matangkasombut, O., Buratowski, R.M., Swilling, N.W., and Buratowski, S. (2000). Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. Genes & development *14*, 951-962.

Matsui, T., Segall, J., Weil, P.A., and Roeder, R.G. (1980). Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. The Journal of biological chemistry *255*, 11992-11996.

Maxon, M.E., Goodrich, J.A., and Tjian, R. (1994). Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance. Genes & development *8*, 515-524.

May, M., Mengus, G., Lavigne, A.C., Chambon, P., and Davidson, I. (1996). Human TAF(II28) promotes transcriptional stimulation by activation function 2 of the retinoid X receptors. The EMBO journal *15*, 3093-3104.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic acids research *41*, W597-600.

Meisterernst, M., and Roeder, R.G. (1991). Family of proteins that interact with TFIID and regulate promoter activity. Cell *67*, 557-567.

Mengus, G., May, M., Jacq, X., Staub, A., Tora, L., Chambon, P., and Davidson, I. (1995). Cloning and characterization of hTAFII18, hTAFII20 and hTAFII28: three subunits of the human transcription factor TFIID. The EMBO journal *14*, 1520-1531.

Merika, M., Williams, A.J., Chen, G., Collins, T., and Thanos, D. (1998). Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. Molecular cell *1*, 277-287.

Mitsiou, D.J., and Stunnenberg, H.G. (2000). TAC, a TBP-sans-TAFs complex containing the unprocessed TFIIAalphabeta precursor and the TFIIAgamma subunit. Molecular cell *6*, 527-537.

Mitsuzawa, H., Seino, H., Yamao, F., and Ishihama, A. (2001). Two WD repeat-containing TATA-binding protein-associated factors in fission yeast that

suppress defects in the anaphase-promoting complex. The Journal of biological chemistry *276*, 17117-17124.

Mizzen, C.A., Yang, X.J., Kokubo, T., Brownell, J.E., Bannister, A.J., Owen-Hughes, T., Workman, J., Wang, L., Berger, S.L., Kouzarides, T., *et al.* (1996). The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. Cell *87*, 1261-1270.

Mohan, W.S., Jr., Scheer, E., Wendling, O., Metzger, D., and Tora, L. (2003). TAF10 (TAF(II)30) is necessary for TFIID stability and early embryogenesis in mice. Molecular and cellular biology *23*, 4307-4318.

Moreland, R.J., Tirode, F., Yan, Q., Conaway, J.W., Egly, J.M., and Conaway, R.C. (1999). A role for the TFIIH XPB DNA helicase in promoter escape by RNA polymerase II. The Journal of biological chemistry *274*, 22127-22130.

Muhlbacher, W., Sainsbury, S., Hemann, M., Hantsche, M., Neyer, S., Herzog, F., and Cramer, P. (2014). Conserved architecture of the core RNA polymerase II initiation complex. Nature communications *5*, 4310.

Muller, F., and Tora, L. (2004). The multicoloured world of promoter recognition complexes. The EMBO journal *23*, 2-8.

Munshi, A., Shafi, G., Aliya, N., and Jyothy, A. (2009). Histone modifications dictate specific biological readouts. Journal of genetics and genomics = Yi chuan xue bao *36*, 75-88.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D.A., Adams, C.M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B.J., *et al.* (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. Science *342*, 1238724.

Muratoglu, S., Georgieva, S., Papai, G., Scheer, E., Enunlu, I., Komonyi, O., Cserpan, I., Lebedeva, L., Nabirochkina, E., Udvardy, A., *et al.* (2003). Two different Drosophila ADA2 homologues are present in distinct GCN5 histone acetyltransferase-containing complexes. Molecular and cellular biology *23*, 306-321.

Narlikar, G.J., Sundaramoorthy, R., and Owen-Hughes, T. (2013). Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. Cell *154*, 490-503.

Nie, Y. (2012). Structural Molecular Biology of Human TFIID Complexes.

Nikolov, D.B., Chen, H., Halay, E.D., Hoffman, A., Roeder, R.G., and Burley, S.K. (1996). Crystal structure of a human TATA box-binding protein/TATA element complex. Proceedings of the National Academy of Sciences of the United States of America *93*, 4862-4867.

Nikolov, D.B., Chen, H., Halay, E.D., Usheva, A.A., Hisatake, K., Lee, D.K., Roeder, R.G., and Burley, S.K. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature *377*, 119-128.

Nurizzo, D., Mairs, T., Guijarro, M., Rey, V., Meyer, J., Fajardo, P., Chavanne, J., Biasci, J.C., McSweeney, S., and Mitchell, E. (2006). The ID23-1 structural biology beamline at the ESRF. Journal of synchrotron radiation *13*, 227-238.

O'Brien, T., and Tjian, R. (1998). Functional analysis of the human TAFII250 N-terminal kinase domain. Molecular cell *1*, 905-911.

Oelgeschlager, T., Chiang, C.M., and Roeder, R.G. (1996). Topology and reorganization of a human TFIID-promoter complex. Nature *382*, 735-738.

Oelgeschlager, T., Tao, Y., Kang, Y.K., and Roeder, R.G. (1998). Transcription activation via enhanced preinitiation complex assembly in a human cell-free system lacking TAFIIs. Molecular cell *1*, 925-931.

Ogryzko, V.V., Kotani, T., Zhang, X., Schiltz, R.L., Howard, T., Yang, X.J., Howard, B.H., Qin, J., and Nakatani, Y. (1998). Histone-like TAFs within the PCAF histone acetylase complex. Cell *94*, 35-44.

Olins, A.L., and Olins, D.E. (1974). Spheroid chromatin units (v bodies). Science *183*, 330-332.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613-622.

Painter, J., and Merritt, E.A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta crystallographica Section D, Biological crystallography *62*, 439-450.

Papai, G., Tripathi, M.K., Ruhlmann, C., Werten, S., Crucifix, C., Weil, P.A., and Schultz, P. (2009). Mapping the initiator binding Taf2 subunit in the structure of hydrated yeast TFIID. Structure *17*, 363-373.

Pardee, T.S., Bangur, C.S., and Ponticelli, A.S. (1998). The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable

RNA polymerase II binding and transcription start site selection. The Journal of biological chemistry *273*, 17859-17864.

Perletti, L., Dantonel, J.C., and Davidson, I. (1999). The TATA-binding protein and its associated factors are differentially expressed in adult mouse tissues. The Journal of biological chemistry *274*, 15301-15304.

Pernot, P., Round, A., Barrett, R., De Maria Antolinos, A., Gobbo, A., Gordon, E., Huet, J., Kieffer, J., Lentini, M., Mattenet, M.*, et al.* (2013). Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. Journal of synchrotron radiation *20*, 660-664.

Peterson, M.G., Tanese, N., Pugh, B.F., and Tjian, R. (1990). Functional domains and upstream activation properties of cloned human TATA binding protein. Science *248*, 1625-1630.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D., Konarev, P.V., and Svergun, D.I. (2012). New developments in the program package for small-angle scattering data analysis. Journal of applied crystallography *45*, 342-350.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry *25*, 1605-1612.

Pham, A.D., and Sauer, F. (2000). Ubiquitin-activating/conjugating activity of TAFII250, a mediator of activation of gene expression in Drosophila. Science *289*, 2357-2360.

Pijnappel, W.W., Esch, D., Baltissen, M.P., Wu, G., Mischerikow, N., Bergsma, A.J., van der Wal, E., Han, D.W., Bruch, H., Moritz, S.*, et al.* (2013). A central role for TFIID in the pluripotent transcription circuitry. Nature *495*, 516-519.

Plaschka, C., Lariviere, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E.V., Borchers, C.H., Baumeister, W., Herzog, F.*, et al.* (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. Nature *518*, 376-380.

Pointud, J.C., Mengus, G., Brancorsini, S., Monaco, L., Parvinen, M., Sassone-Corsi, P., and Davidson, I. (2003). The intracellular localisation of TAF7L, a paralogue of transcription factor TFIID subunit TAF7, is developmentally

regulated during male germ-cell differentiation. Journal of cell science *116*, 1847-1858.

Poon, D., Campbell, A.M., Bai, Y., and Weil, P.A. (1994). Yeast Taf170 is encoded by MOT1 and exists in a TATA box-binding protein (TBP)-TBP-associated factor complex distinct from transcription factor IID. The Journal of biological chemistry *269*, 23135-23140.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nature biotechnology *28*, 1057-1068.

Poss, Z.C., Ebmeier, C.C., and Taatjes, D.J. (2013). The Mediator complex and transcription regulation. Critical reviews in biochemistry and molecular biology *48*, 575-608.

Pugh, B.F., and Tjian, R. (1990). Mechanism of transcriptional activation by Sp1: evidence for coactivators. Cell *61*, 1187-1197.

Rabenstein, M.D., Zhou, S., Lis, J.T., and Tjian, R. (1999). TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. Proceedings of the National Academy of Sciences of the United States of America *96*, 4791-4796.

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J Struct Biol *173*, 530-540.

Robinson, M.M., Yatherajam, G., Ranallo, R.T., Bric, A., Paule, M.R., and Stargell, L.A. (2005). Mapping and functional characterization of the TAF11 interaction with TFIIA. Molecular and cellular biology *25*, 945-957.

Roeder, R.G., and Rutter, W.J. (1970). Specific nucleolar and nucleoplasmic RNA polymerases. Proceedings of the National Academy of Sciences of the United States of America *65*, 675-682.

Romier, C., James, N., Birck, C., Cavarelli, J., Vivares, C., Collart, M.A., and Moras, D. (2007). Crystal structure, biochemical and genetic characterization of yeast and E. cuniculi TAF(II)5 N-terminal domain: implications for TFIID assembly. J Mol Biol *368*, 1292-1306.

Ruppert, S., and Tjian, R. (1995). Human TAFII250 interacts with RAP74: implications for RNA polymerase II initiation. Genes & development *9*, 2747-2755.

Sainsbury, S., Niesser, J., and Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature *493*, 437-440.

Samuels, M., Fire, A., and Sharp, P.A. (1982). Separation and characterization of factors mediating accurate transcription by RNA polymerase II. The Journal of biological chemistry *257*, 14419-14427.

Sanders, S.L., Garbett, K.A., and Weil, P.A. (2002). Molecular characterization of Saccharomyces cerevisiae TFIID. Molecular and cellular biology *22*, 6000-6013.

Saunders, A., Core, L.J., and Lis, J.T. (2006). Breaking barriers to transcription elongation. Nature reviews Molecular cell biology *7*, 557-567.

Schagger, H. (2006). Tricine-SDS-PAGE. Nature protocols *1*, 16-22.

Scheer, E., Delbac, F., Tora, L., Moras, D., and Romier, C. (2012). TFIID TAF6-TAF9 complex formation involves the HEAT repeat-containing C-terminal domain of TAF6 and is modulated by TAF5 protein. The Journal of biological chemistry *287*, 27580-27592.

Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophysical journal *78*, 1606-1619.

Selleck, W., Howley, R., Fang, Q., Podolny, V., Fried, M.G., Buratowski, S., and Tan, S. (2001). A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. Nature structural biology *8*, 695-700.

Shaikh, T.R., Gao, H., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A., and Frank, J. (2008). SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nature protocols *3*, 1941-1974.

Shao, H., Revach, M., Moshonov, S., Tzuman, Y., Gazit, K., Albeck, S., Unger, T., and Dikstein, R. (2005). Core promoter binding by histone-like TAF complexes. Molecular and cellular biology *25*, 206-219.

Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. Journal of biomolecular NMR *44*, 213-223.

Sievers, F., and Higgins, D.G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol *1079*, 105-116.

Sims, R.J., 3rd, Mandal, S.S., and Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. Current opinion in cell biology *16*, 263-271.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annual review of biochemistry *72*, 449-479.

Solow, S., Salunek, M., Ryan, R., and Lieberman, P.M. (2001). Taf(II) 250 phosphorylates human transcription factor IIA on serine residues important for TBP binding and transcription activity. The Journal of biological chemistry *276*, 15886-15892.

Sorzano, C.O., Marabini, R., Velazquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H., Carazo, J.M., and Pascual-Montano, A. (2004). XMIPP: a new generation of an open-source image processing package for electron microscopy. J Struct Biol *148*, 194-204.

Soutoglou, E., Demeny, M.A., Scheer, E., Fienga, G., Sassone-Corsi, P., and Tora, L. (2005). The nuclear import of TAF10 is regulated by one of its three histone fold domain-containing interaction partners. Molecular and cellular biology *25*, 4092-4104.

Stark, H. (2010). GraFix: stabilization of fragile macromolecular complexes for single particle cryo-EM. Methods in enzymology *481*, 109-126.

Stevens, A. (1960). Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli B. Biochemical and biophysical research communications *3*, 92-96.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y.*, et al.* (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome research *11*, 677-684.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. Journal of applied crystallography *28*, 768-773.

Tan, S., Hunziker, Y., Sargent, D.F., and Richmond, T.J. (1996). Crystal structure of a yeast TFIIA/TBP/DNA complex. Nature *381*, 127-151.

Tanese, N., Pugh, B.F., and Tjian, R. (1991). Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. Genes & development *5*, 2212-2224.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: An extensible image processing suite for electron microscopy. J Struct Biol *157*, 38-46.

Taylor, N.M., Baudin, F., von Scheven, G., and Muller, C.W. (2013). RNA polymerase III-specific general transcription factor IIIC contains a heterodimer resembling TFIIF Rap30/Rap74. Nucleic acids research *41*, 9183-9196.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211-1213.

Thomas, M.C., and Chiang, C.M. (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology *41*, 105-178.

Timmers, H.T., Meyers, R.E., and Sharp, P.A. (1992). Composition of transcription factor B-TFIID. Proceedings of the National Academy of Sciences of the United States of America *89*, 8140-8144.

Tora, L. (2002). A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. Genes & development *16*, 673-675.

Trowitzsch, S., Viola, C., Scheer, E., Conic, S., Chavant, V., Fournier, M., Papai, G., Ebong, I.O., Schaffitzel, C., Zou, J.*, et al.* (2015). Cytoplasmic TAF2-TAF8-TAF10 complex provides evidence for nuclear holo-TFIID assembly from preformed submodules. Nature communications *6*, 6011.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J Struct Biol *116*, 17-24.

van Ingen, H., van Schaik, F.M., Wienk, H., Ballering, J., Rehmann, H., Dechesne, A.C., Kruijzer, J.A., Liskamp, R.M., Timmers, H.T., and Boelens, R. (2008). Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. Structure *16*, 1245-1256.

Venters, B.J., and Pugh, B.F. (2009). How eukaryotic genes are transcribed. Critical reviews in biochemistry and molecular biology *44*, 117-141.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell *131*, 58-69.

Verrijzer, C.P., Chen, J.L., Yokomori, K., and Tjian, R. (1995). Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. Cell *81*, 1115-1125.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. Journal of applied crystallography *36*, 860-864.

Voss, N.R., Yoshioka, C.K., Radermacher, M., Potter, C.S., and Carragher, B. (2009). DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. J Struct Biol *166*, 205-213.

Wang, H., Curran, E.C., Hinds, T.R., Wang, E.H., and Zheng, N. (2014). Crystal structure of a TAF1-TAF7 complex in human transcription factor IID reveals a promoter binding module. Cell research *24*, 1433-1444.

Wang, P.J., and Page, D.C. (2002). Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. Human molecular genetics *11*, 2341-2346.

Wang, W., Carey, M., and Gralla, J.D. (1992). Polymerase II promoter activation: closed complex formation and ATP-driven start site opening. Science *255*, 450-453.

Wang, X., Truckses, D.M., Takada, S., Matsumura, T., Tanese, N., and Jacobson, R.H. (2007). Conserved region I of human coactivator TAF4 binds to a short hydrophobic motif present in transcriptional regulators. Proceedings of the National Academy of Sciences of the United States of America *104*, 7839-7844.

Wei, Y., Liu, S., Lausen, J., Woodrell, C., Cho, S., Biris, N., Kobayashi, N., Yokoyama, S., and Werner, M.H. (2007). A TAF4-homology domain from the corepressor ETO is a docking platform for positive and negative regulators of transcription. Nature structural & molecular biology *14*, 653-661.

Weil, P.A., and Blatti, S.P. (1976). HeLa cell deoxyribonucleic acid dependent RNA polymerases: function and properties of the class III enzymes. Biochemistry *15*, 1500-1509.

Weil, P.A., Luse, D.S., Segall, J., and Roeder, R.G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. Cell *18*, 469-484.

Weiss, S.B., and Gladstone, L. (1959). A MAMMALIAN SYSTEM FOR THE INCORPORATION OF CYTIDINE TRIPHOSPHATE INTO RIBONUCLEIC ACID1. Journal of the American Chemical Society *81*, 4118-4119.

Werten, S., Mitschler, A., Romier, C., Gangloff, Y.G., Thuault, S., Davidson, I., and Moras, D. (2002). Crystal structure of a subcomplex of human transcription factor TFIID formed by TATA binding protein-associated factors hTAF4 (hTAF(II)135) and hTAF12 (hTAF(II)20). The Journal of biological chemistry *277*, 45502-45509.

Wieczorek, E., Brand, M., Jacq, X., and Tora, L. (1998). Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. Nature *393*, 187-191.

Wierzbicki, A.T., Ream, T.S., Haag, J.R., and Pikaard, C.S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. Nature genetics *41*, 630-634.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A.*, et al.* (2011). Overview of the CCP4 suite and current developments. Acta crystallographica Section D, Biological crystallography *67*, 235-242.

Wollmann, P., Cui, S., Viswanathan, R., Berninghausen, O., Wells, M.N., Moldt, M., Witte, G., Butryn, A., Wendler, P., Beckmann, R.*, et al.* (2011). Structure and mechanism of the Swi2/Snf2 remodeller Mot1 in complex with its substrate TBP. Nature *475*, 403-407.

Woodcock, C.L., Sweetman, H.E., and Frado, L.L. (1976). Structural repeating units in chromatin. II. Their isolation and partial characterization. Experimental cell research *97*, 111-119.

Wright, K.J., Marr, M.T., 2nd, and Tjian, R. (2006). TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. Proceedings of the National Academy of Sciences of the United States of America *103*, 12347-12352.

Xie, X., Kokubo, T., Cohen, S.L., Mirza, U.A., Hoffmann, A., Chait, B.T., Roeder, R.G., Nakatani, Y., and Burley, S.K. (1996). Structural similarity

between TAFs and the heterotetrameric core of the histone octamer. Nature *380*, 316-322.

Yudkovsky, N., Ranish, J.A., and Hahn, S. (2000). A transcription reinitiation intermediate that is stabilized by activator. Nature *408*, 225-229.

Zhou, J., Zwicker, J., Szymanski, P., Levine, M., and Tjian, R. (1998). TAFII mutations disrupt Dorsal activation in the Drosophila embryo. Proceedings of the National Academy of Sciences of the United States of America *95*, 13483-13488.

Zylber, E.A., and Penman, S. (1971). Products of RNA polymerases in HeLa cell nuclei. Proceedings of the National Academy of Sciences of the United States of America *68*, 2861-2865.